Heedong Goh

Lecture Notes for

Finite Element Method

Department of Civil and Environmental Engineering
Seoul National University
2025

Preface

This monograph is written for the course *Finite Element Method* offered in the Department of Civil and Environmental Engineering at Seoul National University.

In preparing these notes, I borrowed many parts from the following books and lecture notes: [Hughes, 2012, Lee, 2022, Demkowicz, 2023, Engquist, 2014, Cook, 2001, Ciarlet, 2002, Quarteroni et al., 2006].

These lecture notes are a working manuscript, subject to ongoing refinement and enhancement; I welcome and greatly appreciate any reports regarding errors, typos, or any other inaccuracies found within the notes.

Canada Goose at Flat Rock Brook Nature Center, Englewood, NJ, USA

Contents

Chapter 1

Preliminaries

1.1 Index notation

We follow *Einstein notation* and use regular font for both scalars and tensors, whose types are to be inferred from context. For example, a vector v expressed in terms of a basis g_i is written as

$$v = v^i g_i. \tag{1.1}$$

In most cases, we make no distinction between *vectors* and *covectors*, as a Cartesian basis is assumed unless stated otherwise. Accordingly, we also write $v = v^i g_i = v_i g^i = v_i g_i$. Similarly, the *inner product* between two geometric vectors is given by

$$(a, b) \equiv a \cdot \overline{b} = a^i \overline{b_i} = a_i \overline{b_i}, \tag{1.2}$$

where $\cdot$ denotes *(single) contraction* and $\overline{(\,)}$ indicates complex conjugation of the subtended quantity.

1.2 Inner product and induced norm

A *set* equipped with inner product is called an *inner product space*. An inner product is a *bilinear form* (or a *sesquilinear form* for complex vectors) such that

$$(\cdot, \cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{F}. \tag{1.3}$$

Here $\mathbb{F}$ is either a real number $\mathbb{R}$ or a complex number $\mathbb{C}$. Thus, an inner product takes two vectors in $\mathcal{V}$ and returns a scalar.

In addition, an inner product must satisfy the following properties [Oden and Demkowicz, 2017]:

- *Linearity* with respect to the first argument

$$(\alpha u + \beta v, w) = \alpha\,(u, w) + \beta\,(v, w) \quad \forall \alpha, \beta \in \mathbb{F},\ \forall u, v, w \in \mathcal{V}. \tag{1.4}$$

- *Conjugate symmetry*

$$(u, v) = \overline{(v, u)} \quad \forall u, v \in \mathcal{V}. \tag{1.5}$$

- *Positive definiteness*

$$(u,u) > 0 \quad \forall u \neq 0,\ u \in \mathcal{V}. \tag{1.6}$$

Note that an inner product is *anti-linear* with respect to the second argument, i.e.,

$$(u, \alpha v) = \overline{\alpha}\,(u,v) \quad \forall \alpha \in \mathbb{F},\ \forall u,v \in \mathcal{V}. \tag{1.7}$$

Orthogonality is defined such that

$$(u,v) = 0. \tag{1.8}$$

Any inner product space is equipped with an *induced norm* such that

$$\|u\| \equiv \sqrt{(u,u)}. \tag{1.9}$$

Theorem 1.2.1 (Cauchy–Schwarz inequality) *Let $(\cdot,\cdot)$ be an inner product on a vector space $\mathcal{V}$. Then, $\forall u,v \in \mathcal{V}$,*

$$|(u,v)| \leq \|u\|\,\|v\|\,. \tag{1.10}$$

Equality holds when u and v are linearly dependent.

Triangle inequality can be proved from the Cauchy-Schwarz inequality as

$$\begin{aligned} \|u+v\|^2 &= (u+v, u+v) \\ &= (u,u) + (u,v) + (v,u) + (v,v) \\ &= (u,u) + 2\mathrm{Re}\left\{(u,v)\right\} + (v,v) \\ &\leq \|u\|^2 + 2\left|(u,v)\right| + \|v\|^2 \\ &\leq \|u\|^2 + 2\,\|u\|\,\|v\| + \|v\|^2 \\ &= \left(\|u\| + \|v\|\right)^2. \end{aligned} \tag{1.11}$$

Thus, we have $\|u+v\| \leq \|u\| + \|v\|$.

The Cauchy–Schwarz inequality arises in numerous contexts, including the determination of admissible function spaces in weak formulations.

1.3 Function space

In these lecture notes, we consider only *Hilbert spaces*:

- *L^2 space* is an inner product space equipped with L^2 norm, e.g.,

$$\|u\|_{L^2(\Omega)} = \left(\int_\Omega |u|^2\right)^{1/2}. \tag{1.12}$$

- *Sobolev space of the first order* is

$$H^1(\Omega) \equiv \left\{u \in L^2(\Omega)\ :\ \mathrm{grad}\, u \in L^2(\Omega)^3\right\}. \tag{1.13}$$

The corresponding inner product is

$$\begin{aligned}(u,v)_{H^1(\Omega)} &= (u,v) + (\operatorname{grad} u, \operatorname{grad} v) \\ &= \int_\Omega u\overline{v} + \int_\Omega \operatorname{grad} u \cdot \overline{\operatorname{grad} v} \end{aligned} \tag{1.14}$$

and the induced norm is

$$\|u\|_{H^1(\Omega)} = \sqrt{(u,u)_{H^1(\Omega)}}. \tag{1.15}$$

- *Sobolev space of the k-th order* is

  $$H^k(\Omega) \equiv \left\{u \in L^2(\Omega) \ : \ D^\alpha u \in L^2(\Omega), \ \forall |\alpha| \le k\right\}. \tag{1.16}$$

  Here, $D^\alpha = \partial^{|\alpha|}/\left(\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots \partial x_n^{\alpha_n}\right)$ denotes partial derivative of order $|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_n$ in the weak sense.

- H (div) space

  $$H(\operatorname{div}, \Omega) = \left\{u \in L^2(\Omega)^3 \ : \ \operatorname{div} u \in L^2(\Omega)\right\}. \tag{1.17}$$

- H (curl) space

  $$H(\operatorname{curl}, \Omega) = \left\{u \in L^2(\Omega)^3 \ : \ \operatorname{curl} u \in L^2(\Omega)^3\right\}. \tag{1.18}$$

1.4 Smoothness

C^k functions are continuous upto its k-th derivative. For example, $\sin x$ is a C^∞ function.

1.5 Differentiation

The *directional derivative* of a function, or a functional, $u(x)$ with respect to x in the direction v is defined as

$$D_v u \equiv \lim_{\epsilon \to 0} \frac{u(x+\epsilon v) - u(x)}{\epsilon}. \tag{1.19}$$

If the map $v \to D_v u$ is linear, then the directional derivative $D_v u$ is identified as *Gateaux differential.* Then, we have

$$D_v u = (\operatorname{grad} u, v), \tag{1.20}$$

where $\operatorname{grad} u$ is called the gradient of u.

Weak derivative of a generalized function u is defined such that

$$(u, \phi') = -(v, \phi). \tag{1.21}$$

Here, v is the weak derivative of u and ϕ is a test function.

1.6 Integration by parts

Integration by parts in one dimension reads

$$\int_0^1 \frac{du}{dx} v = [uv]_0^1 - \int_0^1 u \frac{dv}{dx}. \tag{1.22}$$

The multi-dimensional generalization of the above equation reads

$$\int_\Omega \frac{\partial u}{\partial x^i} v = \int_{\partial\Omega} u v n_i - \int_\Omega u \frac{\partial v}{\partial x^i}. \tag{1.23}$$

In the above, $\Omega \in \mathbb{R}^N$ is the domain in N dimension and its boundary is denoted by $\partial\Omega$, and n_i is the i-th component of the outward normal vector on the boundary.

Integration by parts for divergence and curl operators are (respectively)

$$\int_\Omega (\operatorname{div} u)\, v = \int_{\partial\Omega} (u \cdot n)\, v - \int_\Omega u \cdot \operatorname{grad} v \quad \text{and} \tag{1.24}$$

$$\int_\Omega \operatorname{curl} E \cdot F = \int_{\partial\Omega} (n \times E) \cdot F + \int_\Omega E \cdot \operatorname{curl} F. \tag{1.25}$$

In the above, u, v, E, and F are vector-valued functions and n is unit outward normal vector. Integration by parts involves with the divergence of a *two tensor*, $A \in \mathbb{R}^{N\times N}$, reads

$$\int_\Omega (\operatorname{div} A) \cdot v = \int_{\partial\Omega} (An) \cdot v - \int_\Omega A : \operatorname{grad} v, \tag{1.26}$$

where : denotes *double contraction* such that $A : B = A_{ij}B^{ij}$.

Chapter 2

Introduction

2.1 Change of basis

The *(contravariant) component* of a vector v^i corresponds to an orthonormal basis e_i can be easily obtained by computing its projection on the basis, i.e.,

$$(v, e_i) = v^i. \tag{2.1}$$

However, we need a rigorous approach for non-orthonormal bases; the contravariant component of a vector v^i is obtained by its projection on the *dual basis* g^i, i.e.,

$$\left(v, g^i\right) = \left(v^j g_j, g^i\right) = v^j \left(g_j, g^i\right) = v^j \delta^i_j = v^i. \tag{2.2}$$

Here, δ^i_j is the *Kronecker delta.* Similarly, we can calculate the *covariant component* of a vector by

$$(v, g_i) = v_i. \tag{2.3}$$

Dual basis g^i is defined such that

$$\left(g_j, g^i\right) = \delta^i_j, \tag{2.4}$$

where $\delta_{ij} = \delta^{ij} = \delta^j_i = \delta^i_j$ denotes Kronecker delta:

$$\delta^i_j = \left\{ \begin{array}{cc} 0 & i \neq j \\ 1 & i = j \end{array} \right. . \tag{2.5}$$

We can still calculate the contravariant components of a vector using basis, however, in a slightly convoluted way, i.e.,

$$\begin{aligned} (v, g_i) &= \left(v^j g_j, g_i\right) \\ &= v^j \left(g_j, g_i\right). \end{aligned} \tag{2.6}$$

In the above, we have a system of equations, where v^j are the unknowns. Thus, we have a matrix equation $K_{ij} u^j = d_i$, where

$$\underbrace{\begin{bmatrix} (g_1, g_1) & (g_1, g_2) & \dots & (g_1, g_N) \\ (g_2, g_1) & (g_2, g_2) & \dots & (g_2, g_N) \\ \vdots & \vdots & \ddots & \vdots \\ (g_N, g_1) & (g_N, g_2) & \dots & (g_N, g_N) \end{bmatrix}}_{=K_{ij}} \underbrace{\begin{pmatrix} v^1 \\ v^2 \\ \vdots \\ v^N \end{pmatrix}}_{=u^j} = \underbrace{\begin{pmatrix} (v, g_i) \\ (v, g_2) \\ \vdots \\ (v, g_N) \end{pmatrix}}_{=d_i}. \tag{2.7}$$

As a special case, consider an orthogonal basis. Then the matrix K_{ij} becomes diagonal. Thus, we have

$$v^i = \frac{(v, g_i)}{(g_i, g_i)}. \tag{2.8}$$

For an orthonormal basis, (2.8) further reduces to (2.1) as $(g_i, g_i) = 1$, because dual basis and basis are identical.

Example 2.1.1 (Coordinate transform) *Calculate the component of a vector $v = 2e_1 + e_2$ with respect to basis $g_1 = e_1$ and $g_2 = e_1 + e_2$.*
Expanding the (2.6), we have

$$v^1 g_1 \cdot g_1 + v^2 g_2 \cdot g_1 = v \cdot g_1 \quad \text{and} \tag{2.9a}$$

$$v^1 g_1 \cdot g_2 + v^2 g_2 \cdot g_2 = v \cdot g_2. \tag{2.9b}$$

In the above, we have $g_1 \cdot g_1 = e_1 \cdot e_1 = 1$, $g_1 \cdot g_2 = g_2 \cdot g_1 = e_1 \cdot (e_1 + e_2) = 1$, $g_2 \cdot g_2 = (e_1 + e_2) \cdot (e_1 + e_2) = 2$, $v \cdot g_1 = (2e_1 + e_2) \cdot e_1 = 2$, and $v \cdot g_2 = (2e_1 + e_2) \cdot (e_1 + e_2) = 3$. Rewriting (2.9) into a matrix equation, we have

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}. \tag{2.10}$$

Then, the solution to the above equation is

$$v^1 = 1, \quad \text{and} \quad v^2 = 1. \tag{2.11}$$

Alternatively, we can obtain the same result using dual bases, i.e., $g^1 = e_1 - e_2$ and $g^2 = e_2$. Then, the equation (2.2) gives

$$v^1 = v \cdot g^1 = (2e_1 + e_2) \cdot (e_1 - e_2) = 1 \quad \text{and} \tag{2.12a}$$

$$v^2 = v \cdot g^2 = (2e_1 + e_2) \cdot e_2 = 1. \tag{2.12b}$$

In summary, we have $v = 2e_1 + e_2 = g_1 + g_2$.

Example 2.1.2 (Fourier series) *Consider an odd function $f(x)$ defined as below*

$$f(x) = x, \quad x \in (-0.5, 0.5). \tag{2.13}$$

Find the component of sinusoidal basis, i.e.,

$$g_n = \sin(n\pi x), \quad n \in \mathbb{Z}_{++}. \tag{2.14}$$

The above sinusoidal basis is orthogonal; thus, we use (2.8), where

$$(g_n, g_n) = \int_{-0.5}^{0.5} \sin(n\pi x) \sin(n\pi x)\, dx = 0.5, \quad \forall n \in \mathbb{Z}_{++} \quad \text{and} \tag{2.15a}$$

$$(f, g_n) = \int_{-0.5}^{0.5} x \sin(n\pi x)\, dx = \frac{2\sin(n\pi/2)}{(n\pi)^2} - \frac{\cos(n\pi/2)}{n\pi}. \tag{2.15b}$$

Then, we have

$$v^n = \frac{(f, g_n)}{(g_n, g_n)} = \frac{\sin(n\pi/2)}{(n\pi)^2} - \frac{\cos(n\pi/2)}{2n\pi}. \tag{2.16}$$

Finally, the function $f(x)$ is *discretized*, or expressed by linear combinations of basis g_n, i.e.,

$$f(x) = \sum_{n=1}^{\infty} \left(\frac{\sin(n\pi/2)}{(n\pi)^2} - \frac{\cos(n\pi/2)}{2n\pi} \right) \sin(n\pi x). \tag{2.17}$$

2.2 Function approximation

What will happen if the choice of basis does not span the function space, e.g., a truncated Fourier series? In such case, we have an approximation of a function. In fact, we obtain the best approximation with respect to the norm induced by the associated inner product. We can verify this by formulating an optimization problem, or least square error problem: given f and g_n, find v^n such that

$$\min \Pi, \quad \Pi = \frac{1}{2} \|v^n g_n - f\|^2. \tag{2.18}$$

In the above, the objective functional Π is proportional to the square of the error, measured by the induced norm, i.e.,

$$\|a\| = \sqrt{(a, a)}. \tag{2.19}$$

The objective functional Π vanishes if and only if $v^n g_n = f$; otherwise, it will always return a positive non-zero number. Then, the minimization problem can be solved by satisfying the first-order optimality condition, i.e.,

$$\begin{aligned} 0 &= \frac{\partial \Pi}{\partial v^n} \\ &= \frac{1}{2} \frac{\partial}{\partial v^n} \left(v^m g_m - f, v^k g_k - f \right) \\ &= (v^m g_m - f, g_n) \\ &= v^m (g_m, g_n) - (g_n, f). \end{aligned} \tag{2.20}$$

Thus, we have recovered the same expression as (2.6).

Example 2.2.1 (Polynomial approximation) *Approximate a function $f(x) = \sin x$, $x \in (-\pi, \pi)$ using polynomials $g_n = x^n$, $n = 1, 2, \ldots, N$, $N = 5$.*

We use (2.6), where

$$
\begin{aligned}
(g_m, g_n) &= \int_{-\pi}^{\pi} x^{m+n} dx = \frac{\pi^{m+n+1} - (-\pi)^{m+n+1}}{m+n+1}, && \text{(2.21a)}\\
(g_1, \sin x) &= 2\pi, && \text{(2.21b)}\\
(g_2, \sin x) &= 0, && \text{(2.21c)}\\
(g_3, \sin x) &= 2\pi\left(\pi^2 - 6\right), && \text{(2.21d)}\\
(g_4, \sin x) &= 0, \quad \text{and} && \text{(2.21e)}\\
(g_5, \sin x) &= 2\pi\left(120 - 20\pi^2 + \pi^4\right). && \text{(2.21f)}
\end{aligned}
$$

Then, we have $v^1 \approx 0.9879$, $v^2 = 0$, $v^3 \approx -0.1553$, $v^4 = 0$, and $v^5 \approx 0.0056$. Figure 2.1 shows the polynomial approximations of function f using $N = 3$ and $N = 5$.

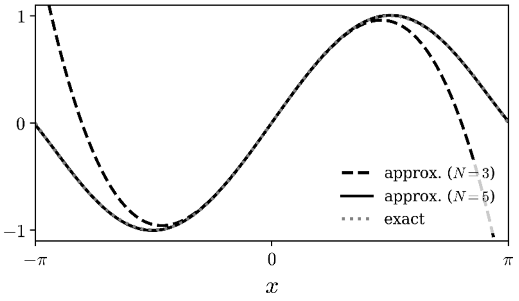

Figure 2.1: Polynomial approximations of $\sin x$ (Example 2.2.1)

Different choices of bases yield different accuracies, computational costs, stability, etc. Finite element method is one example, where we discretize a differential equation, therefore a function/solution, by a set of "local" basis.

2.3 Differential equation and its approximations

A differential equation can be stated abstractly as

$$F\left(u, x, \partial_x^{(n)}, f\right) = 0, \tag{2.22}$$

where $u : \mathbb{F}^d \supset \Omega \to \mathbb{F}^s$ is the unknown function to be determined, x is the coordinates, $\partial_x^{(n)}$ are differential operators, and f are given data such as boundary and initial conditions.

We assume our problems are *wellposed*, i.e., we require

- existence of solution
- uniqueness of solution

- continuous dependence on data: let u and v denote solutions correspond to data f and g, respectively, i.e., $F(u,f) = 0$ and $F(v,g) = 0$. Then, there exists $C \in \mathbb{R}_{++}$ such that

$$\|u - v\|_* \leq C \|f - g\|_{**} . \tag{2.23}$$

In general, the problem $F(u) = 0$ is posed in an infinite-dimensional space. Approximation involves reducing the problem to a finite-dimensional problem $F^h(u^h) = 0$ and solving the resulting system to find u^h, which is a finite-dimensional representation of u. We may consider various representation strategies such as (Figure 2.2) (a) pointwise representation; (b) linear combination of known functions, i.e., $u^h(x) = \alpha_i g_i(x)$; (c) local averages on finite subspaces; and (d) particle distribution, where local density represents the function value. These representations lead to different numerical methods including (a) Finite Difference Method (FDM); (b) Finite Element Method (FEM) (c) Finite Volume Method (FVM); and (d) Particle Method (PM).

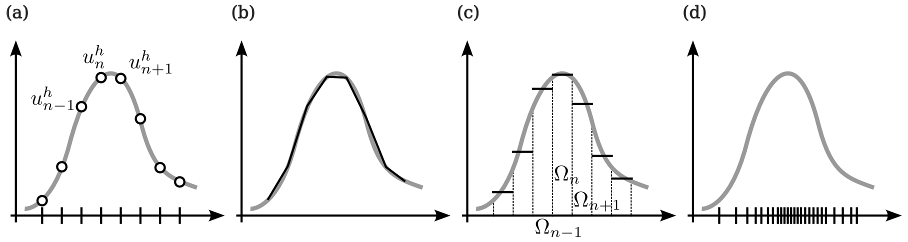

Figure 2.2: Various representation strategies. (a) Pointwise representation. (b) Linear combination of known functions. (c) Local averages. (d) Particle distribution.

Example 2.3.1 (Finite difference method) *Consider a model Poisson's equation:*

$$\begin{cases} \dfrac{d^2u}{dx^2} + f = 0 \quad x \in (0,1) \\ u(0) = u(1) = 0 \end{cases} . \tag{2.24}$$

Use central differencing to approximate, i.e., discretize, the given problem and provide local truncation error.

We approximate a second-order derivative of a function using the central differencing as

$$\frac{d^2u}{dx^2} \approx \frac{u(x-h) - 2u(x) + u(x+h)}{h^2}. \tag{2.25}$$

Denoting $u_{i+1} = u(x_i + h)$, the above governing equation becomes

$$\begin{cases} \dfrac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + f_i = 0 \quad i = 1, 2, \ldots, N-1 \\ u_0 = u_N = 0 \end{cases} , \tag{2.26}$$

where, $h = 1/N$. Rewriting the above equation in a matrix equation, we have

$$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-2} \\ f_{N-1} \end{pmatrix}. \tag{2.27}$$

To calculate the local truncation error, take a Taylor expansion of u_{i+1} about i, which gives

$$\begin{aligned} u_{i+1} &= u\left(x_i + h\right) \\ &\approx u\left(x_i\right) + hu'\left(x_i\right) + \frac{h^2}{2}u''\left(x_i\right) + \frac{h^3}{6}u'''\left(x_i\right) + \frac{h^4}{24}u''''\left(x_i\right) \end{aligned} \tag{2.28}$$

Then, (2.26) becomes

$$\begin{aligned} 0 &= \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + f_i \\ &= \frac{u\left(x_i\right) - hu'\left(x_i\right) + \frac{h^2}{2}u''\left(x_i\right) - \frac{h^3}{6}u'''\left(x_i\right) + \frac{h^4}{24}u''''\left(x_i\right)}{h^2} + \frac{-2u\left(x_i\right)}{h^2} \\ &\quad + \frac{u\left(x_i\right) + hu'\left(x_i\right) + \frac{h^2}{2}u''\left(x_i\right) + \frac{h^3}{6}u'''\left(x_i\right) + \frac{h^4}{24}u''''\left(x_i\right)}{h^2} + f\left(x_i\right) \\ &= u''\left(x_i\right) + f\left(x_i\right) + \frac{h^2}{12}u''''\left(x_i\right) \\ &= \mathcal{O}\left(h^2\right), \end{aligned} \tag{2.29}$$

which reveals the second-order convergence as we refine h.

2.4 Weak form

2.4.1 Model problem

Let us revisit the Poisson's equation:

$$(\mathrm{S})\begin{cases} \dfrac{d^2u}{dx^2} + f = 0 \quad x \in (0,1) \\ u(0) = u(1) = 0 \end{cases}. \tag{2.30}$$

The left-hand side of the governing equation vanishes for all $x \in (0,1)$ when u is a true solution. Thus, the governing equation gives a pointwise requirement of the solution, and such formulation is called a *strong form*.

Alternatively, a *weak form* focuses on the definition of the 0 on the right-hand side; like the left-hand side of the equation, 0 is a function, or a vector. Specifically, the function 0 returns a number 0 when inner-producted with any other function. Thus, we can equivalently write

$$\int v\left(\frac{d^2u}{dx^2}+f\right)=0, \quad \forall v, \tag{2.31}$$

which is identified as a weak form. Here, u is called a *trial function* and the arbitrary function v is called a *test function.*

The above weak form becomes more useful when we take an integration by parts:

$$\begin{aligned} 0 &= \int v\left(\frac{d^2u}{dx^2}+f\right) \\ &= \left[v\frac{du}{dx}\right]_0^1 - \int \frac{dv}{dx}\frac{du}{dx} + \int vf. \end{aligned} \tag{2.32}$$

Here, the boundary term vanishes when the test function satisfies homogeneous Dirichlet boundary conditions, i.e., $v(0)=v(1)=0$. Then, we have

$$\text{(W)}\left\{\begin{array}{l} u\in\mathcal{U} \\ b\left(u,v\right)=l\left(v\right), \quad \forall v\in\mathcal{V} \end{array}\right. . \tag{2.33}$$

In the above, the bilinear and linear forms are defined by

$$b\left(u,v\right)=\int\frac{dv}{dx}\frac{du}{dx} \quad \text{and} \tag{2.34}$$

$$l\left(v\right)=\int vf. \tag{2.35}$$

u and v may be elements of different function spaces $\mathcal{U}$ and $\mathcal{V}$. However, in the given model problem, we derived a symmetric, or Hermitian, bilinear form by integration by parts, which allowed to use the same function space for both trial and test functions:

$$\mathcal{U}=\mathcal{V}=\left\{w\in H^1\left(0,1\right)\ :\ w(0)=w(1)=0\right\}. \tag{2.36}$$

The space $H^1(0,1)$ is chosen to ensure the bilinear form is integrable, i.e.,

$$\left|\left(\frac{du}{dx},\frac{dv}{dx}\right)\right|<\infty, \tag{2.37}$$

where the Cauchy-Schwarz inequality gives

$$\left|\left(\frac{du}{dx},\frac{du}{dx}\right)\right|\le\left\|\frac{du}{dx}\right\|\left\|\frac{du}{dx}\right\|. \tag{2.38}$$

2.4.2 Non-homogeneous boundary conditions

In this section, we discuss a more rigorous approach taking into account non-homogeneous boundary conditions.

The strong form of our model problem reads:
Given $f:\bar{\Omega}\to\mathcal{R}$ and constants p and q, find $u:\bar{\Omega}\to\mathbb{R}$, such that

$$\text{(S)}\left\{\begin{array}{l} \dfrac{d^2u}{dx^2}+f=0, \quad x\in\Omega=(0,1) \\ -\left.\dfrac{du}{dx}\right|_{x=0}=p \\ u|_{x=1}=q \end{array}\right. . \tag{2.39}$$

We multiply the governing equation with a test function v and integrate over the domain.

$$\begin{aligned} 0 &= \int_0^1 v\left(\frac{d^2u}{dx^2}+f\right)dx \\ &= \left[v\frac{du}{dx}\right]_0^1 - \underbrace{\int_0^1 \frac{du}{dx}\frac{dv}{dx}dx}_{=b(u,v)} + \underbrace{\int_0^1 vf\,dx}_{=l(v)}\,. \end{aligned} \tag{2.40}$$

We introduce two function spaces $\mathcal{U}$ and $\mathcal{V}$ such that

$$\mathcal{U}=\mathcal{V}=\left\{u\in H^1(\Omega)\ :\ u(1)=0\right\}. \tag{2.41}$$

Then, we have the following weak form:

$$(\mathrm{W})\left\{\begin{array}{l} u\in\tilde{q}+\mathcal{U} \\ b(u,v)=l(v)+v(0)\,p, \quad \forall v\in\mathcal{V} \end{array}\right. . \tag{2.42}$$

In the above, $\tilde{q}+\mathcal{U}$ denotes the algebraic sum of $\tilde{q}$ and $\mathcal{U}$, i.e.,

$$\tilde{q}+\mathcal{U}\equiv\{\tilde{q}+w\ :\ w\in\mathcal{U}\}. \tag{2.43}$$

Here, $\tilde{q}$ is an extension of q to the domain, which is called a *finite energy lift* of q (Figure 2.3). Introducing the modified linear form

$$l^{\mathrm{mod}}(v)=l(v)+v(0)\,p-b(\tilde{q},v), \tag{2.44}$$

we have an alternative form:

$$(\mathrm{W})\left\{\begin{array}{l} w\in\mathcal{U} \\ b(w,v)=l^{\mathrm{mod}}(v), \quad \forall v\in\mathcal{V} \end{array}\right. . \tag{2.45}$$

In the above, we assumed that $\mathcal{U}$ is a subset of a larger space $\mathcal{X}$ and $\tilde{q}\in\mathcal{X}$.

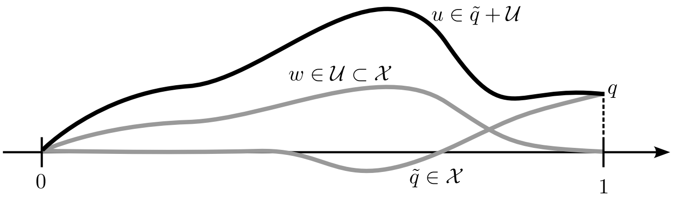
$u\in\tilde{q}+\mathcal{U}$
$w\in\mathcal{U}\subset\mathcal{X}$
q
0
$\tilde{q}\in\mathcal{X}$
1

Figure 2.3: Finite energy lift of q.

2.5 Galerkin method

2.5.1 Galerkin approximation of a weak form

The *Galerkin approximation* replaces the original function spaces to their finite-dimensional subsets: $\mathcal{U}^h\subset\mathcal{U}$ and $\mathcal{V}^h\subset\mathcal{V}$ such that

$$\mathcal{U}^h=\left\{w^h\in\mathcal{U}\ :\ w^h(x)=\sum_{n=1}^N a^n g_n(x)\right\} \quad \text{and} \tag{2.46}$$

$$\mathcal{V}^h=\left\{w^h\in\mathcal{V}\ :\ w^h(x)=\sum_{n=1}^N b^n h_n(x)\right\}. \tag{2.47}$$

In the above, g_n and h_n are known basis functions and a_n and b_n are the coefficients. Then, we have the Galerkin approximation:

$$\text{(G)}\left\{\begin{array}{l} u^h \in \mathcal{U}^h \\ b\left(u^h, v^h\right) = l\left(v^h\right), \quad \forall v^h \in \mathcal{V}^h \end{array}\right. . \tag{2.48}$$

Thus, Galerkin method retains the original problem but approximates the function spaces.

Writing the basis functions explicitly, the Galerkin approximation yields a matrix equation, i.e.,

$$\text{(M)}\left\{\begin{array}{l} a_n \in \mathbb{R} \\ b\left(a_n g_n, b_m h_m\right) = l\left(b_m h_m\right), \quad \forall b_m \in \mathbb{R} \end{array}\right. \tag{2.49}$$

or

$$b\left(g_n, h_m\right) a_n = l\left(h_m\right), \quad m = 1, 2, \ldots, N. \tag{2.50}$$

Here, we have the same form as (2.7).

Example 2.5.1 (Polynomial basis) *Consider a model problem*

$$\left\{\begin{array}{l} \dfrac{d^2u}{dx^2} + 1 = 0, \quad x \in (0,1) \\ u(0) = u(1) = 0 \end{array}\right. \tag{2.51}$$

or, in a weak form,

$$\left\{\begin{array}{l} u^h \in \mathcal{U}^h \\ b\left(u^h, v^h\right) = l\left(v^h\right), \quad \forall v^h \in \mathcal{U}^h \end{array}\right. . \tag{2.52}$$

In the above,

$$b\left(u^h, v^h\right) = \int_0^1 \frac{du^h}{dx}\frac{dv^h}{dx}dx \quad \textit{and} \tag{2.53}$$

$$l\left(v^h\right) = \int_0^1 v^h \cdot 1dx. \tag{2.54}$$

Find u^h when the function space is given by

$$\mathcal{U}^h = \left\{ w^h\left(x\right) = \sum_{n=0}^{2} a_n x^n \ : \ w^h(0) = w^h(1) = 0 \right\}. \tag{2.55}$$

Initially, we have three unknowns, a_0, a_1, and a_2. Applying the boundary condition, we have

$$a_0 = 0 \quad \text{and} \tag{2.56}$$

$$a_0 + a_1 + a_2 = 0. \tag{2.57}$$

Then, we have left with only one unknown, i.e.,

$$u^h = Ax\left(x - 1\right). \tag{2.58}$$

The matrix equation $K_{11}u_1 = f_1$ is obtained by plugging the above function into both trial and test functions in the weak form, where

$$K_{11} = \int_0^1 (2x-1)^2\, dx = \frac{1}{3}, \tag{2.59}$$

$$u_1 = A, \quad \text{and} \tag{2.60}$$

$$f_1 = \int_0^1 x(x-1)\cdot 1 dx = -\frac{1}{6} \tag{2.61}$$

Then, we have $a = -1/2$, which gives

$$u^h = -\frac{1}{2}x(x-1). \tag{2.62}$$

The above solution coincides with the exact solution of the given problem, which occurs when the finite-dimensional subspace happens to contain the true solution.

We can consider a larger subspace, where $u^h = a_0 + a_1x + a_2x^2 + a_3x^3$. When boundary conditions are applied, we have

$$u^h = A_1 \underbrace{x(x-1)}_{g_1} + A_2 \underbrace{x\left(x^2-1\right)}_{g_2}. \tag{2.63}$$

Here, the components of the matrix equation read

$$K_{11} = \int_0^1 (2x-1)^2\, dx = \frac{1}{3}, \tag{2.64}$$

$$K_{12} = K_{21} = \int_0^1 (2x-1)\left(3x^2-1\right) dx = \frac{1}{2}, \tag{2.65}$$

$$K_{22} = \int_0^1 \left(3x^2-1\right)^2 dx = \frac{4}{5}, \tag{2.66}$$

$$f_1 = \int_0^1 x(x-1)\cdot 1 dx = -\frac{1}{6}, \quad \text{and} \tag{2.67}$$

$$f_2 = \int_0^1 x\left(x^2-1\right)\cdot 1 dx = -\frac{1}{4}. \tag{2.68}$$

or

$$\begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & \frac{4}{5} \end{bmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{6} \\ -\frac{1}{4} \end{pmatrix}. \tag{2.69}$$

Then, we recover the same solution since $A_1 = -1/2$ and $A_2 = 0$.

Exercise 2.5.1 (Polynomial basis) *Repeat the above example for*

$$\begin{cases} \dfrac{d^2u}{dx^2} + \pi^2 \sin \pi x = 0, & x \in (0,1) \\ u(0) = u(1) = 0 \end{cases}. \tag{2.70}$$

The function space is given by

$$\mathcal{U}^h = \mathcal{V}^h = \left\{ w^h(x) = \sum_{n=1}^{N} a_n g_n(x) \ : \ g_n(x) = x(x^n - 1) \right\}. \tag{2.71}$$

Try different N and compare the solutions with the exact one:

$$u^{\text{exact}} = \sin \pi x. \tag{2.72}$$

Galerkin method is classified as *Bubnov-Galerkin method* when $\mathcal{U}^h = \mathcal{V}^h$, otherwise is called *Petrov-Galerkin method.* For self-adjoint problems, the natural, but not necessary, choice is the Bubnov-Galerkin method. In many literature, Bubnov-Galerkin method is simply called Galerkin method.

We will mainly discuss self-adjoint elliptic problems, where Galerkin approximation of a weak form, weighted residual method, Galerkin approximation of the variational principle, Ritz method (i.e., minimization of energy over the approximate space), and minimization of function residual become equivalent to each other.

2.5.2 Weighted residual method

The *weighted residual method* starts with the function approximation and identifies the residual as

$$R_E = \frac{d^2 u^h}{dx^2} + f \neq 0. \tag{2.73}$$

Then, we have

$$\begin{aligned} 0 &= \int \phi_m R_E \\ &= \int_0^1 \phi_m \left(\frac{d^2 u^h}{dx^2} + f \right) dx \\ &= \left[\phi_m \frac{du^h}{dx} \right]_0^l - \int_0^1 \frac{d\phi_m}{dx} \frac{du^h}{dx} dx + \int_0^1 \phi_m f dx \\ &= -\, b\left(u^h, \phi_m\right) + l\left(\phi_m\right), \end{aligned} \tag{2.74}$$

which becomes identical with (2.50) when $\phi_m = h_m$ and $u^h = a_n g_n$.

2.5.3 Galerkin approximation of variational principle

We assume a self-adjoint problem. Suppose the energy functional (total potential energy) is given by

$$\Pi[u] = \frac{1}{2} b(u,u) - l(u). \tag{2.75}$$

Here, self-adjoint assumption implies b is symmetric, i.e., $b(u,v) = b(v,u)$ $\forall u, v \in \mathcal{U}$ and, therefore, $\mathcal{U} = \mathcal{V}$.

Following the *principle of minimum potential energy*, we optimize, i.e., use variational principle, the energy functional. The principle of minimum potential energy implies that the solution u is obtained by

$$u = \arg\min_{w \in \mathcal{U}} \Pi\left[w\right]. \tag{2.76}$$

Then, the first-order optimality condition gives

$$\begin{aligned} 0 =& D_u \Pi\left[u\right] \\ =& \frac{1}{2}\left[b\left(u, v\right) + b\left(v, u\right)\right] - l\left(v\right) \\ =& b\left(u, v\right) - l\left(v\right). \end{aligned} \tag{2.77}$$

Then, we approximate trial and test functions using the same basis function, which gives

$$b\left(u^h, v^h\right) - l\left(v^h\right) = 0. \tag{2.78}$$

2.5.4 Ritz method

Ritz method discretize the energy functional of a self-adjoint problem; then, optimize.

Applying the Galerkin approximation of the energy functional, we have

$$\Pi\left[u^h\right] = \frac{1}{2} b\left(u^h, u^h\right) - l\left(u^h\right). \tag{2.79}$$

Then, the first-order optimality condition reads

$$\begin{aligned} 0 =& D_{u^h} \Pi\left[u^h\right] \\ =& \frac{\partial}{\partial a_m} \Pi\left[a_m g_m\right] \\ =& \frac{1}{2}\left[b\left(a_n g_n, b_m g_m\right) + b\left(b_m g_m, a_n g_n\right)\right] - l\left(b_m g_m\right) \\ =& b\left(a_n g_n, b_m g_m\right) - l\left(b_m g_m\right). \end{aligned} \tag{2.80}$$

Here, b_m is the direction of the derivative.

2.5.5 Minimization of function residual

Let us define a *function residual*, or *Galerkin error*, as

$$R_F = u - u^h. \tag{2.81}$$

Then, its minimization in the energy norm reads

$$\left\|R_F\right\|_E = \min_{w^h \in \mathcal{U}^h} \left\|u - w^h\right\|_E, \tag{2.82}$$

where the energy norm is defined as

$$\left\|v\right\|_E^2 = b\left(v, v\right). \tag{2.83}$$

The minimization becomes equivalent with the Ritz method because

$$\begin{aligned}\frac{1}{2}\left\|R_F\right\|_E^2 &= \frac{1}{2}b\left(u-u^h,u-u^h\right)\\ &= \frac{1}{2}b\left(u,u\right)+\frac{1}{2}b\left(u^h,u^h\right)+b\left(u,u^h\right)\\ &= \frac{1}{2}b\left(u,u\right)+\Pi\left[u^h\right].\end{aligned} \tag{2.84}$$

2.5.6 Galerkin orthogonality

Let $u^h \in \mathcal{U}^h \subset \mathcal{U}$ the Galerkin approximation to the variational problem

$$\left\{\begin{array}{l} u^h \in \mathcal{U}^h \subset \mathcal{U} \\ b\left(u^h,v^h\right)=l\left(v^h\right), \quad \forall v^h \in \mathcal{V}^h \subset \mathcal{V} \end{array}\right. . \tag{2.85}$$

Let u denote the exact solution. Subtracting the above problem from

$$b\left(u,v^h\right)=l\left(v^h\right), \quad \forall v^h \in \mathcal{V}^h, \tag{2.86}$$

we obtain the *Galerkin orthogonality*, which states that the error $u-u^h$ satisfies

$$b\left(u-u^h,v^h\right)=0, \quad \forall v^h \in \mathcal{V}^h. \tag{2.87}$$

Chapter 3

One-dimensional Model Problem

3.1 Piecewise linear finite element space

The finite element method is a subclass of the Galerkin method in which the basis functions are locally supported polynomials—nonzero only over a small subdomain. The original domain is partitioned into N nonoverlapping subdomains, called *elements*. Each element contains two or more *nodes*: typically two boundary nodes and, if applicable, additional internal nodes. The basis function defined over an element is also called a *shape function*.

As an example, consider a linear finite element space for a one-dimensional problem, where each element consists of two nodes. Each subdomain, or element, is denoted by $[x_n, x_{n+1}]$, where the length of the element is $h_n = x_{n+1} - x_n$. Then, the function is approximated by

$$u\left(x\right) \approx u^h\left(x\right) = a_n g_n\left(x\right), \tag{3.1}$$

where (Figure 3.1(a))

$$g_n = \begin{cases} \dfrac{x - x_{n-1}}{h_{n-1}}, & x_{n-1} \leq x \leq x_n \\ \dfrac{x_{n+1} - x}{h_n}, & x_n \leq x \leq x_{n+1} \\ 0, & \text{elsewhere} \end{cases} . \tag{3.2}$$

In the above formulation, the coefficients a_n correspond to the nodal values of the function, given that the basis functions are normalized to have unit amplitude at their respective nodes. Note that the basis functions for the boundary nodes are "single-sided", e.g.,

$$g_1 = \begin{cases} \dfrac{x_2 - x}{h_1}, & x_1 \leq x \leq x_2 \\ 0, & \text{elsewhere} \end{cases} . \tag{3.3}$$

The function values within the nodes are linearly interpolated (Figure 3.1(c)).

It is often more convinient to represent the basis functions elementwisely as shown in Figure 3.1(b), i.e.,

$$N_1(\xi) = \frac{\xi_2 - \xi}{h} \quad \text{and} \tag{3.4}$$

$$N_2(\xi) = \frac{\xi - \xi_1}{h}. \tag{3.5}$$

Here, each element has two functions, where its nodal values on the element boundaries are shared across elements.

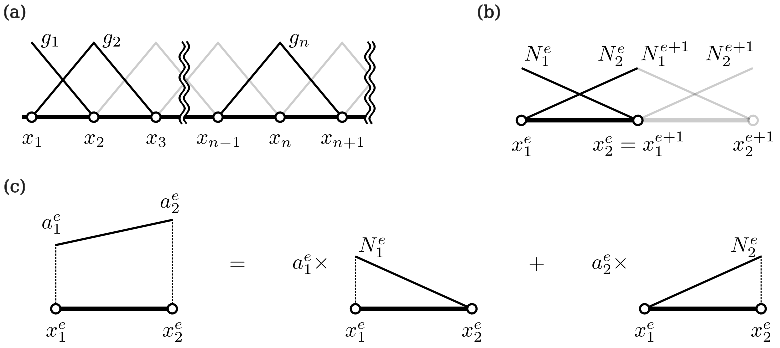

Figure 3.1: Piecewise linear finite element space. (a) linear basis functions. (b) Element-wise representation of the basis function. (c) Linear interpolation.

Note that in the finite element method, the coefficient a_n, also referred to as a *degree-of-freedom*, is identified as the value of u^h at a node x_n by construction. Representing degree-of-freedom as a linear functional ψ_n, we have

$$\left(\psi_n, u^h\right) = u^h(x_n), \tag{3.6}$$

which is a very important property of finite element method.

3.2 Poisson's equation

Consider a model Poisson's equation:

$$\begin{cases} u^h \in \mathcal{U}^h \\ b\left(u^h, v^h\right) = l\left(v^h\right), \quad \forall v^h \in \mathcal{U}^h \end{cases}, \tag{3.7}$$

where

$$b\left(u^h, v^h\right) = \int_0^1 \frac{du^h}{dx}\frac{dv^h}{dx}dx, \tag{3.8}$$

$$l\left(v^h\right) = \int_0^1 v^h \cdot f dx = \int_0^1 v^h \cdot 1 dx, \quad \text{and} \tag{3.9}$$

$$\mathcal{U}^h = \left\{u^h = a_n g_n \ : \ u^h(0) = u^h(1) = 0\right\}. \tag{3.10}$$

Suppose we discretize the domain into N equal-length elements of size $h = 1/N$. Using linear shape functions, we have $N+1$ nodes. Then, the given weak form is written in matrix equation as

$$\left\{ \begin{array}{l} a_n \in \mathbb{R} \\ b\left(a_n g_n, b_m g_m\right) = l\left(b_m g_m\right), \quad \forall b_m \in \mathbb{R} \end{array} \right. . \tag{3.11}$$

Removing dependency in b_m, we have

$$Ka = f \quad \text{or} \quad K_{mn} a_m = f_m, \tag{3.12}$$

where

$$K_{mn} = \int_0^1 \frac{dg_m}{dx}\frac{dg_n}{dx} dx \quad \text{and} \quad f_m = \int_0^1 g_m \cdot 1 dx. \tag{3.13}$$

In the above, K_{mn} is called a *stiffness matrix* and f_m is called a *load vector*. Note that K is symmetric, i.e., $K_{mn} = K_{nm}$, because the bilinear form b is symmetric and (Bubnov-)Garlerkin method is used. Then, the each entry reads

$$\begin{aligned}
K_{00} &= \int_0^h \frac{dg_0}{dx}\frac{dg_0}{dx} dx \\
&= \int_0^h \left(\frac{d}{dx}\frac{h-x}{h}\right)^2 dx = \int_0^h \left(-\frac{1}{h}\right)^2 dx = \frac{1}{h}, && (3.14a) \\
K_{01} &= \int_0^h \frac{dg_0}{dx}\frac{dg_1}{dx} dx \\
&= \int_0^h \left(\frac{d}{dx}\frac{h-x}{h}\right)\left(\frac{d}{dx}\frac{x}{h}\right) dx = \int_0^h \left(-\frac{1}{h}\right)\left(\frac{1}{h}\right) dx \\
&= -\frac{1}{h}, && (3.14b) \\
K_{0n} &= \int_0^h \frac{dg_0}{dx}\frac{dg_n}{dx} dx = 0, \quad n \geq 2, && (3.14c) \\
K_{11} &= \int_0^{2h} \frac{dg_1}{dx}\frac{dg_1}{dx} dx \\
&= \int_0^h \left(\frac{d}{dx}\frac{x}{h}\right)^2 dx + \int_h^{2h} \left(\frac{d}{dx}\frac{2h-x}{h}\right)^2 dx = \frac{2}{h}, && (3.14d) \\
K_{12} &= \int_h^{2h} \frac{dg_1}{dx}\frac{dg_2}{dx} dx \\
&= \int_h^{2h} \left(\frac{d}{dx}\frac{2h-x}{h}\right)\left(\frac{d}{dx}\frac{x-h}{h}\right) dx = -\frac{1}{h}, && (3.14e) \\
\vdots &= \vdots && (3.14f) \\
K_{(N-1)(N-1)} &= \frac{2}{h}, && (3.14g) \\
K_{(N-1)N} &= -\frac{1}{h}, && (3.14h) \\
K_{NN} &= \frac{1}{h}, && (3.14i)
\end{aligned}$$

and

$$f_0 = \int_0^h g_0 \cdot 1dx = \int_0^h \frac{h-x}{h} dx = \frac{h}{2}, \tag{3.15a}$$

$$f_1 = \int_0^{2h} g_1 \cdot 1dx = \int_0^h \frac{x}{h} dx + \int_h^{2h} \frac{2h-x}{h} dx = h, \tag{3.15b}$$

$$\vdots = \vdots \tag{3.15c}$$

$$f_{N-1} = h, \tag{3.15d}$$

$$f_N = \frac{h}{2}. \tag{3.15e}$$

or, in matrix equation,

$$\frac{1}{h}\begin{bmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{bmatrix} \begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} \\ u_N \end{pmatrix} = h \begin{pmatrix} 1/2 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1/2 \end{pmatrix}. \tag{3.16}$$

Since $u_0 = u_N = 0$, the above matrix equation is reduced to

$$\frac{1}{h}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix} = h \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix}. \tag{3.17}$$

The final expression is identical to that of the finite difference method. However, this is a special case, and the two methods differ in general.

3.3 Element-wise formulation

Here, we will derive the same matrix equation in the previous section, however, using an element-wise formulation. While the formulation may appear convoluted, it is actually much more straightforward from a programming perspective and makes it easier to introduce higher-order elements and extend to higher dimensions.

We construct *element stiffness matrices* and *element load vectors.* Then, we *assemble* them to construct a *(global) stiffness matrix* and a *(global) load vector.*

For each element, we construct

$$\tilde{k}^e_{mn} = \int_{x_1^e}^{x_2^e} \frac{dN^e_m}{dx}\frac{dN^e_n}{dx} dx, \ m, n = 1, 2 \quad \text{and} \tag{3.18}$$

$$\tilde{f}^e_m = \int_{x_1^e}^{x_2^e} N^e_m \cdot 1dx, \ m = 1, 2. \tag{3.19}$$

or, assuming every element has the same length h,

$$\tilde{k}^e = \frac{1}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \quad \text{and} \tag{3.20}$$

$$\tilde{f}^e = h\begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} \tag{3.21}$$

such that

$$0 = \sum_{e=1}^{N} b^e \cdot \left(\tilde{k}^e a^e - \tilde{f}^e\right). \tag{3.22}$$

Here, a_1^e and a_2^e denote the coefficients at element e, which are related with the (global) coefficients via compatiblity matrix C^e such that $a^e = C^e a$ or

$$\begin{pmatrix} a_1^e \\ a_2^e \end{pmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 & \cdots & 0 & 0 \end{bmatrix}}_{=C^e} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \\ a_{n+1} \\ \vdots \\ a_{N-1} \\ a_N \end{pmatrix}. \tag{3.23}$$

Similarly, we have $b^e = C^e b$ for the test function. Then, (3.22) becomes

$$\begin{aligned} 0 &= \sum_{e=1}^{N} (C^e b) \cdot \left(\tilde{k}^e C^e a - \tilde{f}^e\right) \\ &= \sum_{e=1}^{N} b \cdot \left((C^e)^T \tilde{k}^e C^e a - (C^e)^T \tilde{f}^e\right) \\ &= b \cdot \left[\sum_{e=1}^{N} \underbrace{\left((C^e)^T \tilde{k}^e C^e\right)}_{=K^e} a - \sum_{e=1}^{N} \underbrace{\left((C^e)^T \tilde{f}^e\right)}_{=f^e}\right] \\ &= b \cdot \left(\underbrace{\sum_{e=1}^{N} K^e}_{=K} a - \underbrace{\sum_{e=1}^{N} f^e}_{=f}\right) \\ &= b \cdot (Ka - f). \end{aligned} \tag{3.24}$$

In some literature, the above assemblabe is denoted by

$$K = \bigcup_{e=1}^{N} \tilde{k}^e \quad \text{and} \tag{3.25}$$

$$f = \bigcup_{e=1}^{N} \tilde{f}^e, \tag{3.26}$$

where $\bigcup$ is called the *assembly operator.*

3.4 Assemblage

The compatibility matrix is mostly sparse, so storing the entire matrix is inefficient in terms of both memory and computation. Sparsity is also characteristic of the global stiffness matrix; however, its treatment will not be addressed in these notes. In numerical implementations, we assign *tags* to nodes and elements to define their connectivity, and assign *IDs* to degrees-of-freedoms (DOF) for determining row and column indices for global matrices and vectors.

Consider a Poisson's problem with non-homogeneous Dirichlet and Neumann boundary conditions:

$$\begin{cases} w^h \in \mathcal{U}^h \\ b\left(w^h, v^h\right) = l^{\text{mod}}\left(v^h\right), \quad \forall v^h \in \mathcal{U}^h \end{cases}, \tag{3.27}$$

where

$$l^{\text{mod}}\left(v^h\right) = l\left(v^h\right) + v^h\left(0\right)p - b\left(\tilde{q}, v^h\right). \tag{3.28}$$

Recall that

$$\mathcal{U}^h = \left\{w^h \in H^1(0,1) \ : \ w^h = a^n g_n, \ w^h(1) = 0\right\}. \tag{3.29}$$

Assume that we have discretized the domain $\Omega = (0,1)$ into 5 subdomains with $|\Omega_e| = h = 1/5$. Figure 3.2 illustrates the domain and tags for nodes and elements.

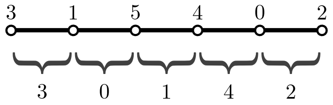

Figure 3.2: Node and element tags. Tags are intentionally unordered.

Then, the connectivity, i.e., node tags for each element, is assigned as Table 3.1.

Table 3.1: Connectivity.

Element tag	0	1	2	3	4
Node tags	(1,5)	(5,4)	(0,2)	(3,1)	(4,0)

Note that we have non-homegeneous Neumann condition on node 3 and non-homogeneous Dirichlet condition on node 2. When assigning IDs for each DOF, we assign lower numbers for unknown DOFs and higher numbers for prescribed DOFs. For example, in Table 3.2, the total number of *free DOFs* is 4 and the number of *constrained DOFs* is 1, where the ID for node 2 is assigned to 5.

Table 3.2: IDs.

Node tag	0	1	2	3	4	5
ID	0	1	5	2	3	4

Given connectivity and ID information, we can assemble element stiffness matrices and vectors. For instance, the element 2 is associated with nodes 0 and 2, where their IDs are 0 and 5, respectively. Namely, the element stiffnes matrix and load vector are assemed in

$$K = \left[\begin{array}{c|c|c|c|c|c} \bullet & & & & & \bullet \\ \hline & & & & & \\ \hline & & & & & \\ \hline & & & & & \\ \hline & & & & & \\ \hline \bullet & & & & & \bullet \end{array}\right] \quad \text{and} \quad f = \left(\begin{array}{c} \bullet \\ \hline \\ \hline \\ \hline \\ \hline \\ \hline \bullet \end{array}\right). \tag{3.30}$$

Let $\tilde{k}^e_{ij}$ denote the (i, j)-th entry of the e-th element stiffness matrix, and let $\tilde{f}^e_i$ denote the i-th entry of the corresponding element load vector. Then, the assembled global system is written as

$$K = \left[\begin{array}{ccccc|c} \tilde{k}^2_{00} + \tilde{k}^4_{11} & 0 & 0 & \tilde{k}^4_{10} & 0 & \tilde{k}^2_{01} \\ 0 & \tilde{k}^0_{00} + \tilde{k}^3_{11} & \tilde{k}^3_{10} & 0 & \tilde{k}^0_{01} & 0 \\ 0 & \tilde{k}^3_{01} & \tilde{k}^3_{00} & 0 & 0 & 0 \\ \tilde{k}^4_{01} & 0 & 0 & \tilde{k}^1_{11} + \tilde{k}^4_{00} & \tilde{k}^1_{10} & 0 \\ 0 & \tilde{k}^0_{10} & 0 & \tilde{k}^1_{01} & \tilde{k}^0_{11} + \tilde{k}^1_{00} & 0 \\ \hline \tilde{k}^2_{10} & 0 & 0 & 0 & 0 & \tilde{k}^2_{11} \end{array}\right] \quad \text{and} \tag{3.31}$$

$$f = \left(\begin{array}{c} \tilde{f}^2_0 + \tilde{f}^4_1 \\ \tilde{f}^0_0 + \tilde{f}^3_1 \\ \tilde{f}^3_0 \\ \tilde{f}^1_1 + \tilde{f}^4_0 \\ \tilde{f}^0_1 + \tilde{f}^1_0 \\ \hline \tilde{f}^2_1 \end{array}\right) \tag{3.32}$$

such that

$$\underbrace{\left[\begin{array}{c|c} K_{ff} & K_{fs} \\ \hline K_{sf} & K_{ss} \end{array}\right]}_{=K} \underbrace{\left(\begin{array}{c} a_f \\ \hline a_s \end{array}\right)}_{=a} = \underbrace{\left(\begin{array}{c} f_f \\ \hline f_s \end{array}\right)}_{=f} + f^{\mathrm{N}} + f^{\mathrm{D}}. \tag{3.33}$$

The above partition separates knowns from unknowns: a_f are the unknown coefficients to be determined, while a_s is prescribed to be zero.

As given in (3.33), the load vector must be supplemented with the non-homogeneous boundary data. The contribution from the non-homogeneous Neumann data $v^h(0)p$ gives

$$f^{\mathrm{N}} = \left(\begin{array}{c} f^{\mathrm{N}}_f \\ \hline f^{\mathrm{N}}_s \end{array}\right) = \left(\begin{array}{c} 0 \\ 0 \\ p \\ 0 \\ 0 \\ \hline 0 \end{array}\right). \tag{3.34}$$

Here, the vector has only a single nonzero entry because the test function evaluated at any node is associated with a single basis function (as implied in (3.6)).

Similarly, while any finite energy lift of q is acceptable, it is convenient to lift q with the same basis that is used for w^h. Namely,

$$\tilde{q}\left(x\right) = c_n g_n\left(x\right) \tag{3.35}$$

such that

$$\tilde{q}\left(x_n\right) = \left\{ \begin{array}{ll} q, & x_n = 1 \\ 0, & \text{otherwise} \end{array} \right., \text{ where } x_n \text{ are nodes.} \tag{3.36}$$

Then, the contribution from the non-homogeneous Dirichlet data $b(\tilde{q}, v)$ yields

$$\begin{aligned} f^{\mathrm{D}} &= \left(\frac{f_f^{\mathrm{D}}}{f_s^{\mathrm{D}}} \right) = - \left[\begin{array}{c|c} K_{ff} & K_{fs} \\ \hline K_{sf} & K_{ss} \end{array} \right] \left(\frac{0}{q_s} \right), \quad \text{where} && (3.37) \\ q_s &= \left(\begin{array}{c} q \end{array} \right). && (3.38) \end{aligned}$$

Finally the reduced equation for solving the unknown vector a_f reads

$$\begin{aligned} K_{ff} a_f &= f_f^{\mathrm{mod}} \\ &= f_f + f_f^{\mathrm{N}} + f_f^{\mathrm{D}} \\ &= f_f + f_f^{\mathrm{N}} - K_{fs} q_s. \end{aligned} \tag{3.39}$$

Exercise 3.4.1 (Poisson's problem with non-homogeneous BC) *Given the strong form:*

$$(\mathrm{S}) \left\{ \begin{array}{ll} \dfrac{d^2u}{dx^2} + u + 1 = 0 & x \in (0,1) \\ u(0) = 1 & \\ \dfrac{du}{dx}(1) = 1 & \end{array} \right. . \tag{3.40}$$

Derive weak form (W), its Galerkin approximation (G), and matrix equation (M) using linear shape functions.

Chapter 4

Linear Elasticity

4.1 Strong form

Here, we review linear elasticity, which consists of

- Balance of momentum

  $$\operatorname{div}\sigma + \rho\omega^2 u + f = 0 \quad \text{or} \quad \sigma_{ij,j} + \rho\omega^2 u_i + f_i = 0. \tag{4.1}$$

  In the above, σ is the stress tensor, ρ is the mass density, ω is the frequency, f is the body force, and u is the displacement.

- Balance of moment of momentum

  $$\sigma = \sigma^T \quad \text{or} \quad \sigma_{ij} = \sigma_{ji}. \tag{4.2}$$

- Constitutive relation

  $$\sigma = C\,[\epsilon] \quad \text{or} \quad \sigma_{ij} = C_{ijkl}\epsilon_{kl}, \tag{4.3}$$

  where C is the elasticity tensor and ϵ is the strain tensor. For linear isotropic medium, we have

  $$\begin{aligned} C\,[\,] &= \mu\,[\,] + \mu\,[\,]^T + \lambda \text{tr}\,[\,]\,I \quad \text{or} \\ C_{ijkl} &= \mu\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{kj}\right) + \lambda\delta_{ij}\delta_{kl}, \end{aligned} \tag{4.4}$$

  where λ and μ are the Lamé parameters and δ_{ij} is Kronecker delta. The above elasticity tensor satisfies the major symmetry

  $$C_{ijkl} = C_{klij} \tag{4.5}$$

  and the minor symmetries

  $$C_{ijkl} = C_{jikl} \quad \text{and} \quad C_{ijkl} = C_{ijlk}. \tag{4.6}$$

- Strain-displacement relation

  $$\epsilon = \frac{1}{2}\left[(\operatorname{grad} u) + (\operatorname{grad} u)^T\right] \quad \text{or} \quad \epsilon_{ij} = \frac{1}{2}\left(u_{i,j} + u_{j,i}\right). \tag{4.7}$$

- Cauchy's theorem
$$t = \sigma n \quad \text{or} \quad t_i = \sigma_{ij} n_j, \tag{4.8}$$
where t is the traction vector and n is the unit outward normal vector.

- Displacement boundary condition
$$u - \bar{u} = 0 \quad \text{or} \quad u_i - \bar{u}_i = 0, \quad x \in \Gamma_u. \tag{4.9}$$
Here, $\bar{u}$ is the prescribed displacement.

- Traction boundary condition
$$t - \bar{t} = 0 \quad \text{or} \quad t_i - \bar{t}_i, \quad x \in \Gamma_t. \tag{4.10}$$
Here, $\bar{t}$ is the prescribed traction.

For the purpose of introducing various weak forms in the subsquent section, we define

- compliance tensor D such that
$$\epsilon = D[\sigma] \quad \text{or} \quad \epsilon_{ij} = D_{ijkl}\sigma_{kl} \quad \text{and} \tag{4.11}$$

- linearized rigid body motion
$$r = \frac{1}{2}\left[(\operatorname{grad} u) - (\operatorname{grad} u)^T\right] \quad \text{or} \quad r_{ij} = \frac{1}{2}\left(u_{i,j} - u_{j,i}\right). \tag{4.12}$$

Then the constitutive relation becomes
$$D[\sigma] = \operatorname{grad} u - r \quad \text{or} \quad D_{ijkl}\sigma_{kl} = u_{i,j} - r_{ij}. \tag{4.13}$$

Given the above notations, we consider the following elastodynamic problem:
$$\begin{cases} -\operatorname{div}\sigma - \rho\omega^2 u = f & \text{in } \Omega \\ D[\sigma] - \operatorname{grad} u + r = 0 & \text{in } \Omega \\ \sigma - \sigma^T & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_u \\ \sigma n = 0 & \text{on } \Gamma_t \end{cases} \tag{4.14}$$
or, in Cartesian,
$$\begin{cases} -\sigma_{ij,j} - \rho\omega^2 u_i = f_i & \text{in } \Omega \\ D_{ijkl}\sigma_{kl} - u_{i,j} + r_{ij} = 0 & \text{in } \Omega \\ \sigma_{ij} - \sigma_{ji} & \text{in } \Omega \\ u_i = 0 & \text{on } \Gamma_u \\ \sigma_{ij} n_j = 0 & \text{on } \Gamma_t \end{cases}. \tag{4.15}$$

4.2 Weak forms

In this section, we provide a brief overview of several weak forms compiled in [Demkowicz, 2023], among many other possible formulations. Similarly to previous examples all weak forms are abstractly notated by
$$\begin{cases} \mathtt{u} \in \mathcal{U} \\ b(\mathtt{u}, \mathtt{v}) = l(\mathtt{v}), \ \forall \mathtt{v} \in \mathcal{V} \end{cases}. \tag{4.16}$$
In the above, `u` and `v` are (group) trial and test functions.

4.2.1 Trivial formulation

In this formulation, we simply multiply the strong form with test functions and integrate over the domain. Both displacement and traction boundary conditions are strongly imposed; therefore, there is no *relaxation.*

The weak form reads

$$\left\{\begin{array}{l} -\left(\operatorname{div}\sigma, v\right) - \left(\rho\omega^2 u, v\right) = \left(f, v\right) \\ \left(D\left[\sigma\right], \tau\right) - \left(\operatorname{grad} u, \tau\right) + \left(r, \tau\right) = 0 \\ \left(\sigma, s\right) = 0 \end{array}\right. . \tag{4.17}$$

In the above, the trial functions, are

$$\mathtt{u} = \left(u, \sigma, r\right). \tag{4.18}$$

Their function spaces are given by

$$u \in H^1\left(\Omega\right)^3 \;:\; u = 0 \text{ on } \Gamma_u, \tag{4.19}$$

$$\sigma \in H\left(\operatorname{div}, \Omega\right)^3 \;:\; \sigma n = 0 \text{ on } \Gamma_t, \quad \text{and} \tag{4.20}$$

$$r = -r^T \in L^2\left(\Omega\right)^3. \tag{4.21}$$

The corresponding test functions are

$$\mathtt{v} = \left(v, \tau, s\right), \tag{4.22}$$

where

$$v \in L^2\left(\Omega\right)^3, \tag{4.23}$$

$$\tau \in L^2\left(\Omega\right)^{3\times 3}, \quad \text{and} \tag{4.24}$$

$$s = -s^T \in L^2\left(\Omega\right)^3. \tag{4.25}$$

In the weak form (4.26), the stress tensor σ is not strongly enforced to be symmetric; instead, symmetry is imposed weakly through the third equation. This formulation is adopted because discretizing the symmetric $H(\operatorname{div})$ space is challenging, whereas discretizing the (anti-)symmetric L^2 space is comparatively straightforward.

Alternatively, letting τ to be symmetric, we have

$$\left\{\begin{array}{l} -\left(\operatorname{div}\sigma, v\right) - \left(\rho\omega^2 u, v\right) = \left(f, v\right) \\ \left(D\left[\sigma\right], \tau\right) - \left(\operatorname{grad} u, \tau\right) = 0 \\ \left(\sigma, s\right) = 0 \end{array}\right. . \tag{4.26}$$

The corresponding function spaces for $\mathtt{u} = (u, \sigma)$ and $\mathtt{v} = (v, \tau, s)$ are

$$u \in H^1\left(\Omega\right)^3 \;:\; u = 0 \text{ on } \Gamma_u, \tag{4.27}$$

$$\sigma \in H\left(\operatorname{div}, \Omega\right)^3 \;:\; \sigma n = 0 \text{ on } \Gamma_t, \tag{4.28}$$

$$v \in L^2\left(\Omega\right)^3, \tag{4.29}$$

$$\tau = \tau^T \in L^2\left(\Omega\right)^6, \quad \text{and} \tag{4.30}$$

$$s = -s^T \in L^2\left(\Omega\right)^3. \tag{4.31}$$

Both formulations (4.17) and (4.26) are non-symmetric; therefore, the Bubnov–Galerkin method cannot be applied.

4.2.2 Relaxed formulation I

In this formulation, we relax the balance of momentum by integrating by parts, which gives

$$\begin{cases} (\sigma, \operatorname{grad} v) - \left(\rho\omega^2 u, v\right) = (f, v) \\ (D[\sigma], \tau) - (\operatorname{grad} u, \tau) + (r, \tau) = 0 \\ (\sigma, s) = 0 \end{cases} . \tag{4.32}$$

In the above, the boundary term arising from integration by parts vanishes, as it is incorporated into the choice of boundary conditions for the function spaces.

The function spaces for $\mathtt{u} = (u, \sigma, r)$ are

$$u \in H^1(\Omega)^3 \ : \ u = 0 \text{ on } \Gamma_u, \tag{4.33}$$

$$\sigma \in L^2(\Omega)^{3\times 3}, \quad \text{and} \tag{4.34}$$

$$r = -r^T \in L^2(\Omega)^3 . \tag{4.35}$$

The function spaces for $\mathtt{v} = (v, \tau, s)$ are

$$v \in H^1(\Omega)^3 \ : \ v = 0 \text{ on } \Gamma_u, \tag{4.36}$$

$$\tau \in L^2(\Omega)^{3\times 3}, \quad \text{and} \tag{4.37}$$

$$s = -s^T \in L^2(\Omega)^3 . \tag{4.38}$$

Similarly, as before, setting $\sigma = \sigma^T$ and $\tau = \tau^T$, (4.32) is reduced to

$$\begin{cases} (\sigma, \operatorname{grad} v) - \left(\rho\omega^2 u, v\right) = (f, v) \\ (D[\sigma], \tau) - (\operatorname{grad} u, \tau) = 0 \end{cases} . \tag{4.39}$$

The corresponding function spaces for $\mathtt{u} = (u, \sigma)$ and $\mathtt{v} = (v, \tau)$ are

$$u \in H^1(\Omega)^3 \ : \ u = 0 \text{ on } \Gamma_u, \tag{4.40}$$

$$\sigma = \sigma^T \in L^2(\Omega)^6, \tag{4.41}$$

$$v \in H^1(\Omega)^3 \ : \ v = 0 \text{ on } \Gamma_u, \quad \text{and} \tag{4.42}$$

$$\tau = \tau^T \in L^2(\Omega)^6 . \tag{4.43}$$

Both formulations (4.32) and (4.39) are symmetric.

4.2.3 Reduced relaxed formulation I

Here, we replace σ in the first equation of (4.39) by constitutive relation, which gives the principle of virtual work.

$$(C[\operatorname{grad} u], \operatorname{grad} v) - \left(\rho\omega^2 u, v\right) = (f, v) . \tag{4.44}$$

The corresponding function spaces are

$$u \in H^1(\Omega)^3 \ : \ u = 0 \text{ on } \Gamma_u, \quad \text{and} \tag{4.45}$$

$$v \in H^1(\Omega)^3 \ : \ v = 0 \text{ on } \Gamma_u. \tag{4.46}$$

This formulation has the smallest number of unknowns.

4.2.4 Relaxed formulation II

Here, we keep the balance of momentum as it were, and relax the constitutive relation, which gives

$$\begin{cases} -\left(\operatorname{div}\sigma, v\right) - \left(\rho\omega^2 u, v\right) = (f, v) \\ \left(D\left[\sigma\right], \tau\right) + \left(u, \operatorname{div}\tau\right) + (r, \tau) = 0 \\ (\sigma, s) = 0 \end{cases} . \tag{4.47}$$

Then, the function spaces for $\mathtt{u} = (u, \sigma, r)$ are

$$u \in L^2\left(\Omega\right)^3, \tag{4.48}$$

$$\sigma \in H\left(\operatorname{div}, \Omega\right)^3 \ : \ \sigma n = 0 \text{ on } \Gamma_t, \quad \text{and} \tag{4.49}$$

$$r = -r^T \in L^2\left(\Omega\right)^3. \tag{4.50}$$

The function spaces for $\mathtt{v} = (v, \tau, s)$ are

$$v \in L^2\left(\Omega\right)^3, \tag{4.51}$$

$$\tau \in H\left(\operatorname{div}, \Omega\right)^3 \ : \ \tau n = 0 \text{ on } \Gamma_t, \quad \text{and} \tag{4.52}$$

$$s = -s^T \in L^2\left(\Omega\right)^3. \tag{4.53}$$

Here, we have a symmetric function space setting.

4.2.5 Reduced relaxed formulation II

Assuming $\omega \neq 0$, we replace u in the relaxed constitutive relation using the first equation, i.e., balance of momentum, of (4.47). Then, we have

$$\begin{cases} \left(D\left[\sigma\right], \tau\right) - \left(\omega^{-2}\rho^{-1}\operatorname{div}\sigma, \operatorname{div}\tau\right) + (r, \tau) = \left(\omega^{-2}\rho^{-1}f, \operatorname{div}\tau\right) \\ (\sigma, s) = 0 \end{cases} . \tag{4.54}$$

The corresponding function spaces for $\mathtt{u} = (\sigma, r)$ and $\mathtt{v} = (\tau, s)$ are

$$\sigma \in H\left(\operatorname{div}, \Omega\right)^3 \ : \ \sigma n = 0 \text{ on } \Gamma_t, \tag{4.55}$$

$$r = -r^T \in L^2\left(\Omega\right)^3, \tag{4.56}$$

$$\tau \in H\left(\operatorname{div}, \Omega\right)^3 \ : \ \tau n = 0 \text{ on } \Gamma_t, \quad \text{and} \tag{4.57}$$

$$s = -s^T \in L^2\left(\Omega\right)^3. \tag{4.58}$$

4.2.6 Ultra weak formulation

Here, we relax both balance of momentum and constitutive relation:

$$\begin{cases} \left(\sigma, \operatorname{grad} v\right) - \left(\rho\omega^2 u, v\right) = (f, v) \\ \left(D\left[\sigma\right], \tau\right) + \left(u, \operatorname{div}\tau\right) + (r, \tau) = 0 \\ (\sigma, s) = 0 \end{cases} . \tag{4.59}$$

Then, the function spaces for $\mathtt{u} = (u, \sigma, r)$ are

$$u \in L^2\left(\Omega\right)^3, \tag{4.60}$$

$$\sigma \in L^2\left(\Omega\right)^{3\times 3}, \quad \text{and} \tag{4.61}$$

$$r = -r^T \in L^2\left(\Omega\right)^3. \tag{4.62}$$

The function spaces for $\mathtt{v} = (v, \tau, s)$ are

$$v \in H^1(\Omega)^3 \ : \ v = 0 \text{ on } \Gamma_u, \tag{4.63}$$

$$\tau \in H(\operatorname{div}, \Omega)^3 \ : \ \tau n = 0 \text{ on } \Gamma_t, \quad \text{and} \tag{4.64}$$

$$s = -s^T \in L^2(\Omega)^3 . \tag{4.65}$$

In addition, we may enforce symmetry of σ; then (4.59) yields

$$\begin{cases} (\sigma, \operatorname{grad} v) - \left(\rho\omega^2 u, v\right) = (f, v) \\ (D[\sigma], \tau) + (u, \operatorname{div} \tau) + (r, \tau) = 0 \end{cases} . \tag{4.66}$$

The function spaces for $\mathtt{u} = (u, \sigma, r)$ are

$$u \in L^2(\Omega)^3 , \tag{4.67}$$

$$\sigma = \sigma^T \in L^2(\Omega)^6 , \quad \text{and} \tag{4.68}$$

$$r = -r^T \in L^2(\Omega)^3 . \tag{4.69}$$

The function spaces for $\mathtt{v} = (v, \tau)$ are

$$v \in H^1(\Omega)^3 \ : \ v = 0 \text{ on } \Gamma_u, \quad \text{and} \tag{4.70}$$

$$\tau \in H(\operatorname{div}, \Omega)^3 \ : \ \tau n = 0 \text{ on } \Gamma_t. \tag{4.71}$$

Exercise 4.2.1 (Symmetric or not) *Verify for each weak form in this chapter whether the bilinear form $b(\mathtt{u}, \mathtt{v})$ is symmetric or asymmetric.*

4.3 Coercivity

Coercive (but not necessarily symmetric) problems are relatively easy to ensure existence, uniqueness, and stability of solutions by Lax-Milgram Theorem and Céa's Lemma. Consider a weak form with a symmetric functional setting:

$$\begin{cases} u \in \mathcal{U} \\ b(u, v) = l(v), \quad \forall v \in \mathcal{U} \end{cases} , \tag{4.72}$$

We say that the given sesquilinear form is *$\mathcal{U}$-coercive* when there exists a constant $\alpha > 0$ such that

$$\alpha \|u\|_{\mathcal{U}}^2 \leq b(u, u), \quad \forall u \in \mathcal{U}. \tag{4.73}$$

Theorem 4.3.1 (Lax-Milgram Theorem) *Let $\mathcal{U}$ be a Hilbert space, let $b(u, v)$ be a continuous and coercive sesquilinear form defined on $\mathcal{U} \times \mathcal{U}$, and let $l(v)$ be a continuous anti-linear form. Then the abstract variational problem*

$$\begin{cases} u \in \mathcal{U} \\ b(u, v) = l(v), \quad \forall v \in \mathcal{U} \end{cases} \tag{4.74}$$

is then well-posed, i.e., it admits a unique solution that depends continuously upon the data.

Theorem 4.3.2 (Céa's Lemma) *Let $b(u,v)$ be a continuous, or bounded, and coercive sesquilinear form defined on Hilbert space U, i.e.,*

$$|b(u,v)| \leq M\|u\|\|v\|, \quad u,v \in \mathcal{U} \quad \textit{(continuity)} \quad \textit{and} \tag{4.75}$$

$$|b(u,u)| \geq \alpha\|u\|^2, \quad u \in \mathcal{U},\ \alpha > 0 \quad \textit{(coercivity).} \tag{4.76}$$

Let $\mathcal{U}^h \subset \mathcal{U}$, and let $u^h \in \mathcal{U}^h$ be the Bubnov-Galerkin projection of some $u \in \mathcal{U}$ onto subspace $\mathcal{U}^h$, i.e.,

$$b\left(u-u^h, v^h\right) = 0, \quad \forall v^h \in \mathcal{U}^h. \tag{4.77}$$

Then, the following stability result holds:

$$\underbrace{\left\|u-u^h\right\|_{\mathcal{U}}}_{\text{approximation error}} \leq \frac{M}{\alpha} \underbrace{\inf_{w^h \in \mathcal{U}^h} \left\|u-w^h\right\|_{\mathcal{U}}}_{\text{the best approximation error}}. \tag{4.78}$$

Here, M/α is called the *stability constant.*

The above theorem can be proved by using coercivity, Galerkin orthogonality, and continuity such that

$$\begin{aligned}
\alpha\left\|u-u^h\right\|_{\mathcal{U}}^2 &\leq \left|b\left(u-u^h, u-u^h\right)\right| \\
&= \left|b\left(u-u^h, u-w^h+w^h-u^h\right)\right| \\
&= \left|b\left(u-u^h, u-w^h\right) + b\left(u-u^h, w^h-u^h\right)\right| \\
&= \left|b\left(u-u^h, u-w^h\right)\right| \\
&\leq M\left\|u-u^h\right\|_{\mathcal{U}}\left\|u-w^h\right\|_{\mathcal{U}},
\end{aligned} \tag{4.79}$$

which gives

$$\left\|u-u^h\right\|_{\mathcal{U}} \leq \frac{M}{\alpha} \inf_{w^h \in \mathcal{U}^h} \left\|u-w^h\right\|_{\mathcal{U}}. \tag{4.80}$$

As discussed in Section 2.5, a symmetric and coercive weak form is equivalent to the minimization problem in the energy norm. Consequently, the stability constant with respect to the energy norm equals one, and the weak form yields the orthogonal projection, i.e., the best approximation error.

4.3.1 Elastostatics

As an example, we consider an *elastostatic* problem, i.e., $\omega = 0$, with a symmetric functional setting. For example, the principle of virtual work gives

$$b(u,v) = (C[\operatorname{grad} u], \operatorname{grad} v) \quad \text{and} \tag{4.81}$$

$$l(v) = (f,v). \tag{4.82}$$

In this context, coercivity implies positive-definiteness of the stored energy.

The elasticity tensor C_{ijkl} is *uniformly, or strictily, elliptic*, i.e.,

$$C_{ijkl}A_{ij}A_{kl} \geq a_0 A_{ij}A_{ij}, \quad a_0 > 0,\ \forall A_{ij} = A_{ji}. \tag{4.83}$$

Then, we have

$$(C\,[\operatorname{grad} u]\,, \operatorname{grad} u) = (C\,[\epsilon]\,, \epsilon) \geq a_0\,(\epsilon, \epsilon) = a_0 \sum_{i,j} \|\epsilon_{ij}\|^2_{L^2(\Omega)}\,. \tag{4.84}$$

Now consider the two theorems:

Theorem 4.3.3 (Poincaré inequality) *Let Ω be a bounded domain in $\mathbb{R}^N$, and let Γ_1 is a subset of $\partial\Omega$ with a positive measure. There exists a positive constant $a_1 > 0$ such that*

$$a_1\,\|u\|^2_{H^1(\Omega)} \leq \|\operatorname{grad} u\|^2_{L^2(\Omega)}\,, \quad \forall u \in H^1\,(\Omega)\;:\; u = 0 \text{ on } \Gamma_1. \tag{4.85}$$

and

Theorem 4.3.4 (Korn's inequality) *Let Ω be a bounded domain in $\mathbb{R}^N$, and let Γ_1 is a subset of $\partial\Omega$ with a positive measure. There exists a positive constant $a_2 > 0$ such that*

$$a_2\,\|\operatorname{grad} u\|^2_{L^2(\Omega)} \leq \sum_{i,j} \|\epsilon_{ij}\|^2_{L^2(\Omega)}\,, \quad \forall u \in H^1\,(\Omega)^N\;:\; u = 0 \text{ on } \Gamma_1. \tag{4.86}$$

Then, we have the coercivity:

$$\begin{aligned}(C\,[\operatorname{grad} u]\,, \operatorname{grad} u) &\geq a_0 \sum_{i,j} \|\epsilon_{ij}\|^2_{L^2(\Omega)} \\ &\geq a_2 a_0\,\|\operatorname{grad} u\|^2_{L^2(\Omega)} \\ &\geq a_2 a_1 a_0\,\|u\|^2_{H^1(\Omega)}\,.\end{aligned} \tag{4.87}$$

Exercise 4.3.1 (Bar problem) *Consider a bar problem*

$$\begin{cases} \dfrac{d}{dx}\left[EA\dfrac{du}{dx}\right] + f = 0, & x \in (0,1) \\ u(0) = u(1) = 0 \end{cases}. \tag{4.88}$$

Let $EA\,(x) \geq k > 0$ is strictly positive definite in $(0,1)$. Prove the coercivity of the symmetric bilinear form:

$$b\,(u,v) = \int_0^1 \frac{du}{dx} EA \frac{dv}{dx} dx. \tag{4.89}$$

Explain how the coercivity property is reflected in the stiffness matrix of the discrete system when the Bubnov–Galerkin method is applied.

Chapter 5

Electromagnetism

5.1 Strong form

The time-harmonic ($e^{i\omega t}$) form of Maxwell's equations is expressed as

$$\left\{\begin{array}{ll} \operatorname{curl} E = -J_m^{\mathrm{imp}} - i\omega B & \text{(Faraday's law)} \\ \operatorname{curl} H = J^{\mathrm{imp}} + i\omega D & \text{(Maxwell-Ampère's law)} \\ \operatorname{div} D = \rho & \text{(Gauss's law)} \\ \operatorname{div} B = \rho_m & \text{(Gauss's magnetic law)} \end{array}\right. . \tag{5.1}$$

Here, E and H denote the electric and magnetic fields, respectively; D is the electric flux density (or electric displacement field), and B is the magnetic flux density. The prescribed impressed electric current and charge density are denoted by J^{imp} and ρ, while their hypothetical magnetic counterparts are represented by J_m^{imp} and ρ_m. Physically, we have $J_m^{\mathrm{imp}} = 0$ and $\rho_m = 0$; however, their inclusion can be useful in computational electromagnetism for maintaining formal symmetry.

For a linear medium, the constitutive relation reads

$$\left\{\begin{array}{l} D = \varepsilon E \\ B = \mu H \end{array}\right. , \tag{5.2}$$

where ε and μ are permittivity and permeability, respectively.

For boundary conditions, we have

- prescribed magnetic surface current

$$n \times E - \underbrace{n \times \bar{E}}_{=-J_m^{\mathrm{S,imp}}} = 0, \quad \text{on } \Gamma_E. \tag{5.3}$$

Here, $J_m^{\mathrm{S,imp}}$ denotes a prescribed magnetic surface current. The special case $n \times \bar{E} = 0$ corresponds to a perfect electric conductor (PEC).

- prescribed electric surface current

$$n \times H - \underbrace{n \times \bar{H}}_{=J^{\mathrm{S,imp}}} = 0, \quad \text{on } \Gamma_H. \tag{5.4}$$

Here, $J^{\mathrm{S,imp}}$ represents a prescribed electric surface current. A hypothetical condition $n \times \bar{H} = 0$ corresponds to a perfect magnetic conductor (PMC).

- impedance boundary condition

$$n \times H + dE_t - J^{\mathrm{S,imp}} = 0, \quad \text{on } \Gamma_i, \tag{5.5}$$

where $E_t = -n \times (n \times E)$ is the tangential component of E and d is a prescribed impedance.

Using the vector triple product identity,

$$a \times (b \times c) = b\,(a \cdot c) - c\,(a \cdot b)\,,$$

we have

$$\begin{aligned} -n \times (n \times E) &= E - n\,(n \cdot E) \\ &= E_t. \end{aligned}$$

Note that $n \times E$ is also a tangential vector because

$$(n \times E) \cdot n = 0,$$

which is rotated by $\pi/2$.

5.2 Weak forms

TODO: TBA

Chapter 6

Shape Functions

6.1 Basic properties of finite elements

Finite element method is a kind of Galerkin method with a specific process of constructing the subspace $\mathcal{U}^h$. In [Ciarlet, 2002], the basic properties of the finite element space is defined by

1. Triangulation $\mathcal{T}^h$ is established over the set $\bar{\Omega}$, i.e.,

   $$\bar{\Omega} = \bigcup_{K \in \mathcal{T}^h} K, \tag{6.1}$$

   where the interior of elements K have no overlap:

   $$K_i^{\circ} \cap K_j^{\circ} = \emptyset, \quad \forall i \neq j. \tag{6.2}$$

2. A space of finite element shape functions $X(K)$ for each $K \in \mathcal{T}^h$ contains polynomials or “nearly polynomials”.

3. (Unisolvence condition) Degrees-of-freedom (DOFs) ψ_j form a basis in the algebraic dual of $X(K)$, i.e.,

   $$(\psi_j, \phi_i) = \delta_{ij}, \quad \phi_i \in X(K). \tag{6.3}$$

   Here, ϕ_i are identified as finite element shape functions. Thus, given the values of the DOFs, i.e., the coefficients, a function is uniquely interpolated.

We define element interpolation operator as

$$\Pi_K u \equiv \sum_j^n (\psi_j, u)\, \phi_j \in X(K). \tag{6.4}$$

6.2 H^1-conforming Lagrange elements

H^1-conforming elements are those for which the finite element space is a subset of $H^1(\Omega)$. This holds if and only if the space is globally continuous across element boundaries.

Here, we consider the *Lagrange element*, which enforces C^0 continuity across the element boundaries. For a Lagrange element, the element degrees-of-freedom are defined as:

$$\psi_j \;:\; X\left(K\right) \ni u \to u\left(a_j\right) \in \mathbb{R}, \tag{6.5}$$

where, a_j are the Lagrange nodes.

6.2.1 Courant's triangle

The Courant triangle refers to a Lagrange triangle of order $p = 1$, where the element shape function space is the polynomial space of degree one:

$$X\left(K\right) = \mathcal{P}^1\left(K\right) = \operatorname{span}\left\{1, x, y\right\}. \tag{6.6}$$

Namely, we intend to interpolate a function u using linear shape functions such that

$$\Pi_K u\left(x, y\right) = \alpha_0 + \alpha_1 x + \alpha_2 y, \tag{6.7}$$

where α_i are constants to be determined. Let (x_i, y_i), $i = 0, 1, 2$ denote the coordinates of the vertices; then, the unisolvence condition gives

$$\begin{aligned} u_0 &= \alpha_0 + \alpha_1 x_0 + \alpha_2 y_0, \quad && \tag{6.8}\\ u_1 &= \alpha_0 + \alpha_1 x_1 + \alpha_2 y_1, \quad && \text{and} \tag{6.9}\\ u_2 &= \alpha_0 + \alpha_1 x_2 + \alpha_2 y_2. && \tag{6.10} \end{aligned}$$

Here, u_i, $i = 0, 1, 2$ are the nodal values at (x_i, y_i). Thus, α_i are uniquely identified.

Let three vertices are $(0, 0)$, $(1, 0)$, and $(0, 1)$; then, we have $\alpha_0 = u_0$, $\alpha_1 = -u_0 + u_1$, and $\alpha_2 = -u_0 + u_2$, which gives

$$\begin{aligned} \Pi_K u\left(x, y\right) &= u_0 + \left(-u_0 + u_1\right) x + \left(-u_0 + u_2\right) y \\ &= \underbrace{\left(1 - x - y\right)}_{=\phi_0} u_0 + \underbrace{x}_{=\phi_1} u_1 + \underbrace{y}_{=\phi_2} u_2. \end{aligned} \tag{6.11}$$

TODO: figures showing Lagrange nodes and shape functions

6.2.2 Lagrange triangle of order p

A higher-order Lagrange triangular element is constructed by introducing additional nodes, with their total number equal to the dimension of the element's shape-function space. For example, for $p = 3$, the element has 10 nodes, and its element shape function space is given by

$$X\left(K\right) = \mathcal{P}^p\left(K\right) = \operatorname{span}\left\{1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3\right\}. \tag{6.12}$$

The corresponding monomials for an arbitrary polynomial order p can be systematically arranged following Pascal's triangle, as shown in Figure 6.1.

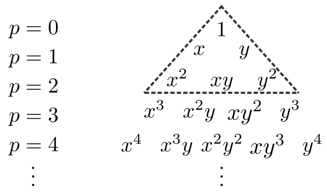

Figure 6.1: Pascal's triangle. Dashed lines indicates monomials for Lagrange triangle of order $p = 2$.

Let $\left(x^{(j)}, y^{(j)}\right)$ are the Lagrange nodes. Then, the ith shape function reads

$$\phi_i = \prod_{j=0\,:\,i\neq j}^{N-1} \frac{x - x^{(j)}}{x^{(i)} - x^{(j)}} \frac{y - y^{(j)}}{y^{(i)} - y^{(j)}}, \quad N = \frac{(p+1)(p+2)}{2}. \tag{6.13}$$

Here $\prod$ is the product operator, and N is the number of Lagrange nodes.

TODO: figures showing Lagrange nodes and shape functions

6.2.3 Isoparametric element

We may use a certain geometry as a master element to construct master shape functions. Then, shape functions for an arbitrary triangle can be derive via a map from the master element. Let $\hat{K}$ denote the master element. We define a map x_K from $\hat{K}$ onto a physical element K, i.e.,

$$x_K \ : \ \hat{K} \ni \xi \to x = x_K(\xi) \in K. \tag{6.14}$$

Thus, the space of elment shape functions are

$$X(K) \equiv \left\{ \hat{u} \circ x_K^{-1} \ : \ \hat{u} \in X\left(\hat{K}\right) \right\}. \tag{6.15}$$

We also have

$$(\psi_j, u) = \left(\hat{\psi}_j, \hat{u}\right), \tag{6.16}$$

where $\hat{\psi}_j$ are the degrees-of-freedom defined in $\hat{K}$ and $u = \hat{u} \circ x_K^{-1}$.

Suppose x_K lives in the master element space of shape functions such that

$$x_K = \sum_j x_{K,j} \hat{\phi}_j(\xi). \tag{6.17}$$

Here, $x_{K,j}$ are the coordinates of the physical element K and $\hat{\phi}_j$ are the shape functions in the master element $\hat{K}$. Then, we identify the that we use *isoparametric finite element*. Returning to the example of the Courant's triangle, we have

$$x = \underbrace{(1-\xi-\eta)}_{=\hat{\phi}_0} x_0 + \underbrace{\xi}_{=\hat{\phi}_1} x_1 + \underbrace{\eta}_{=\hat{\phi}_2} x_2 \quad \text{and} \tag{6.18}$$

$$y = \underbrace{(1-\xi-\eta)}_{=\hat{\phi}_0} y_0 + \underbrace{\xi}_{=\hat{\phi}_1} y_1 + \underbrace{\eta}_{=\hat{\phi}_2} y_2. \tag{6.19}$$

Note that if the element map belongs to a subspace of the shape function space, the element is called *sub-parametric.* If the element map belongs to a superspace of the shape function space, the element is called *super-parametric.*

The derivatives of the shape functions are derived via the chain rule:

$$\frac{\partial \hat{\phi}_i}{\partial \xi_a} = \frac{\partial \phi_i}{\partial x_b}\frac{\partial x_b}{\partial \xi_a} \quad \text{or} \quad \begin{pmatrix} \dfrac{\partial \hat{\phi}_i}{\partial \xi} \\ \dfrac{\partial \hat{\phi}_i}{\partial \eta} \end{pmatrix} = \underbrace{\begin{bmatrix} \dfrac{\partial x}{\partial \xi} & \dfrac{\partial y}{\partial \xi} \\ \dfrac{\partial x}{\partial \eta} & \dfrac{\partial y}{\partial \eta} \end{bmatrix}}_{=J^T} \begin{pmatrix} \dfrac{\partial \phi_i}{\partial x} \\ \dfrac{\partial \phi_i}{\partial y} \end{pmatrix}. \tag{6.20}$$

Thus, we have

$$\operatorname{grad}_x \phi_i = J^{-T} \operatorname{grad}_\xi \hat{\phi}_i. \tag{6.21}$$

Similarly, one can derive

$$d\Omega = (\det J)\, d\hat{\Omega}. \tag{6.22}$$

Then, the Piola transformations read

$$u = \hat{u} \circ x_K^{-1} \quad \text{and} \tag{6.23}$$

$$\operatorname{grad}_x u = J^{-T} \operatorname{grad}_\xi \hat{u} \circ x_K^{-1}. \tag{6.24}$$

TODO: a figure comparing affine map and isoparametric map for a quadratic triangle.

Note that the finite element space must be capable of representing rigid body motions to ensure convergence toward correct results upon h-refinement [Cook, 2001]. The isoparametric formulation inherently satisfies this requirement by guaranteeing the inclusion of rigid body motion [Demkowicz, 2023].

6.2.4 Q4 element

The finite elment space of shape functions for quadrangles is denoted by $\mathcal{Q}^{p,q}$, which is constructed as the tensor product of two one-dimensional polynomial spaces:

$$\mathcal{Q}^{p,q} = \mathcal{P}^p \otimes \mathcal{P}^q. \tag{6.25}$$

For example, Q4 element is constructed by

$$\begin{aligned}\mathcal{Q}^{1,1} &= \mathcal{P}^1 \otimes \mathcal{P}^1 \\ &= \{1, \xi_1\} \otimes \{1, \xi_2\} \\ &= \{1, \xi_1, \xi_2, \xi_1\xi_2\} . \end{aligned} \tag{6.26}$$

Thus, the corresponding shape functions are not purely linear due to the cross-term $\xi_1\xi_2$.

Given the Lagrange nodes located at

$$(\xi_1, \xi_2) = (0,0),\ (1,0),\ (1,1),\ \text{and}\ (0,1), \tag{6.27}$$

the shape functions are:

$$\hat{\phi}_0 = (1 - \xi_1)(1 - \xi_2), \tag{6.28a}$$

$$\hat{\phi}_1 = \xi_1 (1 - \xi_2), \tag{6.28b}$$

$$\hat{\phi}_2 = \xi_1\xi_2, \quad \text{and} \tag{6.28c}$$

$$\hat{\phi}_3 = (1 - \xi_1)\xi_2. \tag{6.28d}$$

TODO: figures showing Lagrange nodes and shape functions

6.2.5 Q9 element

Shape functions of a quadrangle of $\mathcal{Q}^{p,p}$ are given by

$$\hat{\phi}_i = \mu_a^{\xi_1} \mu_b^{\xi_2}, \quad a, b = 0, 1, 2, \ldots, p, \tag{6.29}$$

where

$$\mu_i^{\xi} = \prod_{j=0\,:\,i \neq j}^{p} \frac{\xi - \xi^{(j)}}{\xi^{(i)} - \xi^{(j)}}. \tag{6.30}$$

Here, $\xi^{(i)}$ are coordinates of Lagrange nodes.

TODO: Pascal's triangle

For example, the Lagrange quadrangle of $\mathcal{Q}^{2,2}$ are called $Q9$ elements. In addition to the vertices of $Q4$ element, we define interior Lagrange nodes with in the edges and the element. We have biquadratic shape functions:

- vertex shape functions $(0,0)$, $(1,0)$, $(1,1)$, $(0,1)$:

$$\hat{\phi}_1 = \mu_0^{\xi_1} \mu_0^{\xi_2} \tag{6.31a}$$

$$\hat{\phi}_2 = \mu_1^{\xi_1} \mu_0^{\xi_2} \tag{6.31b}$$

$$\hat{\phi}_3 = \mu_0^{\xi_1} \mu_1^{\xi_2} \tag{6.31c}$$

$$\hat{\phi}_4 = \mu_1^{\xi_1} \mu_1^{\xi_2} \tag{6.31d}$$

- edge bubbles $(1/2, 0)$, $(1, 1/2)$, $(1/2, 1)$, $(0, 1/2)$:

$$\hat{\phi}_5 = \mu_2^{\xi_1} \mu_0^{\xi_2} \tag{6.31e}$$

$$\hat{\phi}_6 = \mu_1^{\xi_1} \mu_2^{\xi_2} \tag{6.31f}$$

$$\hat{\phi}_7 = \mu_2^{\xi_1} \mu_1^{\xi_2} \tag{6.31g}$$

$$\hat{\phi}_8 = \mu_0^{\xi_1} \mu_2^{\xi_2} \tag{6.31h}$$

- element bubble $(1/2, 1/2)$:

$$\hat{\phi}_9 = \mu_2^{\xi_1} \mu_2^{\xi_2} \tag{6.31i}$$

In the above,

$$\mu_0^{\xi} = 2\left(\xi - \frac{1}{2}\right)(\xi - 1), \tag{6.32a}$$

$$\mu_1^{\xi} = 2\xi\left(\xi - \frac{1}{2}\right), \quad \text{and} \tag{6.32b}$$

$$\mu_2^{\xi} = 4\xi(1 - \xi). \tag{6.32c}$$

TODO: figures showing Lagrange nodes and shape functions

6.3 Gauss quadrature

Gauss quadrature can be viewed as a higher-order generalization of the Riemann sum for approximating the integral of a function. While the rectangle rule approximates the function as piecewise constant and the trapezoidal rule as piecewise linear, Gauss quadrature achieves higher accuracy by using higher-order polynomials via optimally selecting integration points and weights.

The general form of the Gauss quadrature is

$$\int_{-1}^{1} f(\xi)\, d\xi \approx \sum_{i=1}^{N} f(\xi_i)\, w_i. \tag{6.33}$$

Here, the right-hand side represents the numerical approximation of the integral, where ξ_i and w_i denote the Gauss points and Gauss weights, respectively. For instance, in the trapezoidal rule, the integration points and weights are given by $(\xi_i, w_i) = (-1, 1/2)$, $(1, 1/2)$.

For example, let us try to determine (ξ_i, w_i) that computes the exact integral for a constant function $f(\xi) = c_0$. Here, a one-point rule with $\xi = 0$ and $w = 2$ is sufficient such that

$$\int_{-1}^{1} c_0 dx = 2c_0 = f(0) \cdot 2. \tag{6.34}$$

Next, we derive a two-point rule for a linear function $f\left(\xi\right) = c_0 + c_1 x$.

$$\int_{-1}^{1} f\left(\xi\right) = 2c_0 = \left(c_0 + c_1\xi_1\right) w_1 + \left(c_0 + c_1\xi_2\right) w_2. \tag{6.35}$$

A simple and effective choice is $\xi_1 = 0$ and $w_2 = 0$, which reduces to the one-point rule derived previously.

We now consider a quadratic function $f\left(\xi\right) = c_0 + c_1 x + c_2 x^2$. Then,

$$\int_{-1}^{1} f\left(\xi\right) = 2c_0 + \frac{2}{3}c_2 = \left(c_0 + c_1\xi_1 + c_2\xi_1^2\right) w_1 + \left(c_0 + c_1\xi_2 + c_2\xi_2^2\right) w_2. \tag{6.36}$$

As in the previous cases, the choice of Gauss points and weights is not unique. Imposing symmetry, i.e., $\xi_1 = -\xi_2$ and $w_1 = w_2$, gives

$$2c_0 + \frac{2}{3}c_2 = \left(2c_0 + 2c_1\xi_1^2\right) w_1, \tag{6.37}$$

from which we obtain the two-point rule:

$$\left(\xi_i, w_i\right) = \left(-\frac{1}{\sqrt{3}}, 1\right), \left(\frac{1}{\sqrt{3}}, 1\right). \tag{6.38}$$

Gauss-Legendre quadrature is generally used, where the Gauss points are the zeros of the Legendre polynomials. N-point Gauss-Legendre rule exactly computes the integral of polynomials of order $2N - 1$.

TODO: a table of points and weights.

Changing the interval from $[-1, 1]$ to $[a, b]$ can be done by

$$\int_{a}^{b} f\left(x\right) dx = \int_{-1}^{1} f\left(x\left(\xi\right)\right) \frac{dx}{d\xi} d\xi, \tag{6.39}$$

where

$$x\left(\xi\right) = \frac{b-a}{2}\xi + \frac{b+a}{2} \quad \text{and} \quad \frac{dx}{d\xi} = \frac{b-a}{2}. \tag{6.40}$$

Gauss quadrature for higher dimensions can be obtained by tensor product of a one-dimensional rule. For example, Gauss quadrature for a quadrangle reads

$$\int_{-1}^{1}\int_{-1}^{1} f\left(\xi, \eta\right) d\xi d\eta \approx \sum_{i=1}^{N}\sum_{j=1}^{N} f\left(\xi_i, \eta_j\right) w_i w_j. \tag{6.41}$$

The quadrature rule for a triangular domain can be obtained through the following transformation, which maps $(\xi, \eta) \in [-1, 1]^2$ to $(r, s) \in T$, where T is a triangle with vertices located at $(0, 0)$, $(1, 0)$, and $(0, 1)$.

$$r = \frac{1+\xi}{2} \quad \text{and} \quad s = \frac{\left(1-\xi\right)\left(1+\eta\right)}{4}. \tag{6.42}$$

The corresponding Jacobian is

$$J = \begin{bmatrix} \frac{\partial r}{\partial \xi} & \frac{\partial r}{\partial \eta} \\ \frac{\partial s}{\partial \xi} & \frac{\partial s}{\partial \eta} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ -\frac{1+\eta}{4} & \frac{1-\xi}{4} \end{bmatrix}, \tag{6.43}$$

whose determinant is

$$\det J = \frac{1-\xi}{8}. \tag{6.44}$$

Thus, the quadrature rule for integration over the triangular domain becomes

$$\int_T f(r,s)\,drds \approx \sum_{i=1}^{N}\sum_{j=1}^{N} f\left(r\left(\xi_i,\eta_j\right), s\left(\xi_i,\eta_j\right)\right) w_i w_j \frac{1-\xi_i}{8}. \tag{6.45}$$

TODO: figures showing GQ for quadrangles and triangles

6.4 Conformity

Previously, we stated that an H^1-conforming finite element must be globally continuous. A more rigorous understanding of this continuity requirement necessitates careful consideration of derivatives in the weak sense.

For example, let $v \in C_0^\infty(\Omega)$ be a test function, where $\Omega \subset \mathbb{R}^3$. Then, the integration by parts on gradients reads

$$-\int_\Omega u \operatorname{grad} v = \int_\Omega v \operatorname{grad} u - \int_{\partial\Omega} unv. \tag{6.46}$$

Let K denote an element in $\bar{\Omega} = \bigcup_{K\in\mathcal{T}^h} K$, and $u_{|K} \in H^1(K)$. The above integration by parts yields

$$-\int_\Omega u \operatorname{grad} v = \sum_K \int_K v \operatorname{grad} u_{|K} - \sum_K \int_{\partial K} u_{|K} n_{|K} v. \tag{6.47}$$

At an interface between adjacent elements, the outward unit normal vectors on each element, denoted by $n_{|K}$, point in opposite directions. Denoting the jump of u across the interface Γ by $[u]_\Gamma$, we have

$$-\int_\Omega u \operatorname{grad} v = \sum_K \int_K v \operatorname{grad} u_{|K} + \sum_\Gamma \int_\Gamma [u]_\Gamma \, nv. \tag{6.48}$$

Recall the definition of weak derivative

$$-(u, \phi') = (v, \phi), \tag{6.49}$$

where v is the weak derivative of u and ϕ is a test function. Then, the gradient of u reads

$$\operatorname{grad} u = \sum_K \operatorname{grad} u_{|K} + \sum_\Gamma [u]_\Gamma \, n\delta_\Gamma. \tag{6.50}$$

Here, δ_Γ is a surface Dirac delta.

Note that $\operatorname{grad} u_{|K}$ belongs to $L^2(K)$, whereas $[u]_\Gamma \, n\delta_\Gamma$ does not, due to the presence of the Dirac delta function. Therefore, to ensure H^1 conformity, we require

$$[u]_\Gamma = 0, \quad \forall \Gamma, \tag{6.51}$$

which corresponds to global C^0 continuity.

Similarly, the continuity requirement for $H(\operatorname{div}, \Omega)$ space is derived from the integration by parts:

$$-\int_\Omega \sigma : \operatorname{grad} v = \int_\Omega v \cdot \operatorname{div} \sigma - \int_{\partial\Omega} \sigma n \cdot v. \tag{6.52}$$

Applying the above formula for $\sigma_{|K} \in H(\operatorname{div}, K)$, we have

$$\begin{aligned} -\int_\Omega \sigma : \operatorname{grad} v &= \sum_K \int_K v \cdot \operatorname{div} \sigma_{|K} + \sum_\Gamma \int_\Gamma [\sigma n]_\Gamma \cdot v \\ &= \sum_K \int_K v \cdot \operatorname{div} \sigma_{|K} + \sum_\Gamma \int_\Omega [\sigma n]_\Gamma \, \delta_\Gamma \cdot v. \end{aligned} \tag{6.53}$$

Then, the divergence reads

$$\operatorname{div} \sigma = \sum_K \operatorname{div} \sigma_{|K} + \sum_\Gamma [\sigma n]_\Gamma \, \delta_\Gamma, \tag{6.54}$$

which gives

$$[\sigma n]_\Gamma = 0, \quad \forall \Gamma, \tag{6.55}$$

Thus, the normal component of σ must be continuous.

The continuity requirement for $H(\operatorname{curl}, \Omega)$ space is derived from

$$\int_\Omega E \cdot \operatorname{curl} F = \int_\Omega F \cdot \operatorname{curl} E - \int_{\partial\Omega} (n \times E) \cdot F. \tag{6.56}$$

For $E_{|K} \in H(\operatorname{curl}, K)$, we have

$$\begin{aligned} \int_\Omega E \cdot \operatorname{curl} F &= \int_\Omega F \cdot \operatorname{curl} E - \int_{\partial\Omega} (n \times E) \cdot F \\ &= \sum_K \int_K F \cdot \operatorname{curl} E_{|K} + \sum_\Gamma \int_\Gamma [n \times E]_\Gamma \cdot F, \end{aligned} \tag{6.57}$$

which implies

$$[n \times E]_\Gamma = 0, \quad \forall \Gamma. \tag{6.58}$$

Thus, the tangential component of E must be continuous.

6.5 Multi-dimensional H^1 space

In elasticity, the typical unknown is the displacement field, which is a three-dimensional vector function. When using the principle of virtual work, both trial and test functions belong to the space $H^1(\Omega)^3$.

Each component of the displacement field can be interpolated independently using an H^1-conforming element. Consequently, each Lagrange node possesses three DOF. The interpolation within an elmenent K can be expressed as

$$u_{i|K} = N_{ij} a_{j|K} \tag{6.59}$$

or

$$\underbrace{\begin{pmatrix} u_{0|K} \\ u_{1|K} \\ u_{2|K} \end{pmatrix}}_{=u_{i|K}} = \underbrace{\begin{bmatrix} \phi_0 & \phi_1 & \dots & \phi_{N-1} & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \phi_0 & \phi_1 & \dots & \phi_{N-1} & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \phi_0 & \phi_1 & \dots & \phi_{N-1} \end{bmatrix}}_{=N_{ij}} \underbrace{\begin{pmatrix} a_{0|K} \\ a_{1|K} \\ a_{2|K} \\ \vdots \\ a_{3N-1|K} \end{pmatrix}}_{=a_{j|K}} . \tag{6.60}$$

Here, the same set of shape functions is used to interpolate each component of the displacement field. The arrangement of N_{ij} may vary depending on the chosen ordering of degrees-of-freedom within the element. In the present example, the horizontal DOFs are assigned first, followed by the vertical DOFs (Figure 6.2).

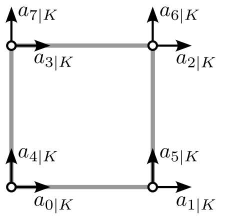

Figure 6.2: Example of element DOF ordering for a Q4 element.

Example 6.5.1 (Elastostatics) *Write an algorithm to construct an element stiffness matrix for elastostatics using an isoparametric element and Gauss–Legendre quadrature.*

The corresponding bilinear form is

$$b(u, v) = \int_\Omega v_{i,j} C_{ijkl} u_{k,l}, \tag{6.61}$$

where

$$\begin{aligned} A_{ij}C_{ijkl}B_{kl} &= A_{ij}\left[\mu\left(\delta_{ik}\delta_{jl}+\delta_{il}\delta_{kj}\right)+\lambda\delta_{ij}\delta_{kl}\right]B_{kl} \\ &= \mu A_{ij}\delta_{ik}\delta_{jl}B_{kl}+\mu A_{ij}\delta_{il}\delta_{kj}B_{kl}+\lambda A_{ij}\delta_{ij}\delta_{kl}B_{kl} \\ &= \mu A_{ij}B_{ij}+\mu A_{ij}B_{ji}+\lambda A_{ii}B_{kk}. \end{aligned} \tag{6.62}$$

For example, in two dimension, the above reduces to

$$\begin{aligned} A_{ij}C_{ijkl}B_{kl} =& \,\mu\left(A_{11}B_{11}+A_{12}B_{12}+A_{21}B_{21}+A_{22}B_{22}\right) \\ &+\mu\left(A_{11}B_{11}+A_{12}B_{21}+A_{21}B_{12}+A_{22}B_{22}\right) \\ &+\lambda\left(A_{11}+A_{22}\right)\left(B_{11}+B_{22}\right). \end{aligned} \tag{6.63}$$

We follow the definition (6.60), where $v_{i,j} = N_{i\alpha,j}b_\alpha$ and $u_{k,l} = N_{k\beta,l}a_\beta$. Then, element stiffness matrix reads

$$\begin{aligned} K^e_{\alpha\beta} &= \int_K N_{i\alpha,j}C_{ijkl}N_{k\beta,l} \\ &= \int_{\hat K} \underbrace{J^{-1}_{aj}\hat N_{i\alpha,a}}_{=A_{ij}(\alpha)}\, C_{ijkl}\, \underbrace{J^{-1}_{bl}\hat N_{k\beta,b}}_{=B_{kl}(\beta)} \det J. \end{aligned} \tag{6.64}$$

We partition the DOF vector a such that

$$a^T = \left(\begin{array}{c|c|c} a^T_{(1)} & a^T_{(2)} & a^T_{(3)} \end{array}\right). \tag{6.65}$$

Here, $a_{(i)} = \left(a_{i\cdot N+0},\, a_{i\cdot N+1},\, \ldots,\, a_{i\cdot N+N-1}\right)^T$ contains DOFs correspond to u_i. Then, we partition the element stiffness matrix as

$$K^e = \left[\begin{array}{c|c|c} k_{(11)} & k_{(12)} & k_{(13)} \\ \hline k_{(21)} & k_{(22)} & k_{(23)} \\ \hline k_{(31)} & k_{(32)} & k_{(33)} \end{array}\right]. \tag{6.66}$$

Frome (6.60), we have $N_{1j} = 0$ for $j \geq N$, $N_{2j} = 0$ for $j < N$ or $j \geq 2N$, and $N_{3j} = 0$ for $j < 2N$. Thus, computation of each $k_{(mn)}$ involves with A_{ij} and B_{kl} constructions with a structured zero pattern:

$$A_{ij}\left(\alpha\right) = J^{-1}_{aj}\hat\phi_{\alpha,a}\delta_{mi} \quad \text{and} \quad B_{kl}\left(\beta\right) = J^{-1}_{bl}\hat\phi_{\beta,b}\delta_{nk}. \tag{6.67}$$

Here, δ_{ij} is Kronecker delta.

For example,

$$k_{(11)} \;:\; A_{ij} = \left[\begin{array}{c|c|c} J^{-1}_{aj}\hat\phi_{\alpha,a} & 0 & 0 \end{array}\right]^T,\; B_{ij} = \left[\begin{array}{c|c|c} J^{-1}_{bl}\hat\phi_{\beta,b} & 0 & 0 \end{array}\right]^T, \tag{6.68a}$$

$$k_{(12)} \;:\; A_{ij} = \left[\begin{array}{c|c|c} J^{-1}_{aj}\hat\phi_{\alpha,a} & 0 & 0 \end{array}\right]^T,\; B_{ij} = \left[\begin{array}{c|c|c} 0 & J^{-1}_{bl}\hat\phi_{\beta,b} & 0 \end{array}\right]^T, \tag{6.68b}$$

$$k_{(13)} \;:\; A_{ij} = \left[\begin{array}{c|c|c} J^{-1}_{aj}\hat\phi_{\alpha,a} & 0 & 0 \end{array}\right]^T,\; B_{ij} = \left[\begin{array}{c|c|c} 0 & 0 & J^{-1}_{bl}\hat\phi_{\beta,b} \end{array}\right]^T, \tag{6.68c}$$

$$k_{(21)} \;:\; A_{ij} = \left[\begin{array}{c|c|c} 0 & J^{-1}_{aj}\hat\phi_{\alpha,a} & 0 \end{array}\right]^T,\; B_{ij} = \left[\begin{array}{c|c|c} J^{-1}_{bl}\hat\phi_{\beta,b} & 0 & 0 \end{array}\right]^T, \tag{6.68d}$$

$$\vdots$$

Note that, for given α and β, $J^{-1}_{aj}\hat{\phi}_{\alpha,a}$ and $J^{-1}_{bl}\hat{\phi}_{\beta,b}$ can be reused for all $k_{(mn)}$ calculations.

For two dimension, see Algorithm 1.

Algorithm 1 Element stiffness matrix for 2D elasticity

```
Initialize $2N \times 2N$ memory space for $k^e$.        ▷ $N$ : number of nodes
for $i = 1, 2, \ldots, N_{\text{gp}}$ do                  ▷ $N_{\text{gp}}$ : number of Gauss points
    Select Gauss points and weights: $\xi_i$, $\eta_i$, $w_i$
    Compute arrays of shape functions and their derivatives at $\xi_i$ and $\eta_i$:

        $\hat{\phi}_a\Big|_{\xi=\xi_i,\eta=\eta_i}, \quad \frac{\partial\hat{\phi}_a}{\partial\xi}\Big|_{\xi=\xi_i,\eta=\eta_i}, \quad \frac{\partial\hat{\phi}_a}{\partial\eta}\Big|_{\xi=\xi_i,\eta=\eta_i}, \quad a = 1, 2, \ldots, N.$

    Compute $x$, $y$, $J^{-T}$, and $\det J$ at $\xi_i$ and $\eta_i$.
    for $\alpha = 1, 2, \ldots, N$ do
        Compute $\phi_\alpha$, $\partial\phi_\alpha/\partial x$, and $\partial\phi_\alpha/y$
        for $\beta = 1, 2, \ldots, N$ do
            Compute $\phi_\beta$, $\partial\phi_\beta/\partial x$, and $\partial\phi_\beta/y$
            • Compute $\psi = A_{ab}C_{abcd}B_{cd}$ with

                $A_{11} = \partial\phi_\alpha/\partial x,\ A_{12} = \partial\phi_\alpha/\partial y,\ A_{21} = 0,\ A_{22} = 0,$
                $B_{11} = \partial\phi_\beta/\partial x,\ B_{12} = \partial\phi_\beta/\partial y,\ B_{21} = 0,\ B_{22} = 0$

            Update $k^e_{\alpha,\beta} \leftarrow k^e_{\alpha,\beta} + \psi \cdot w_i \cdot \det J$
            • Compute $\psi = A_{ab}C_{abcd}B_{cd}$ with

                $A_{11} = \partial\phi_\alpha/\partial x,\ A_{12} = \partial\phi_\alpha/\partial y,\ A_{21} = 0,\ A_{22} = 0,$
                $B_{11} = 0,\ B_{12} = 0,\ B_{21} = \partial\phi_\beta/\partial x,\ B_{22} = \partial\phi_\beta/\partial y$

            Update $k^e_{\alpha,\beta+N} \leftarrow k^e_{\alpha,\beta+N} + \psi \cdot w_i \cdot \det J$
            • Compute $\psi = A_{ab}C_{abcd}B_{cd}$ with

                $A_{11} = 0,\ A_{12} = 0,\ A_{21} = \partial\phi_\alpha/\partial x,\ A_{22} = \partial\phi_\alpha/\partial y,$
                $B_{11} = \partial\phi_\beta/\partial x,\ B_{12} = \partial\phi_\beta/\partial y,\ B_{21} = 0,\ B_{22} = 0$

            Update $k^e_{\alpha+N,\beta} \leftarrow k^e_{\alpha+N,\beta} + \psi \cdot w_i \cdot \det J$ a
            • Compute $\psi = A_{ab}C_{abcd}B_{cd}$ with

                $A_{11} = 0,\ A_{12} = 0,\ A_{21} = \partial\phi_\alpha/\partial x,\ A_{22} = \partial\phi_\alpha/\partial y,$
                $B_{11} = 0,\ B_{12} = 0,\ B_{21} = \partial\phi_\beta/\partial x,\ B_{22} = \partial\phi_\beta/\partial y$

            Update $k^e_{\alpha+N,\beta+N} \leftarrow k^e_{\alpha+N,\beta+N} + \psi \cdot w_i \cdot \det J$
        end for
    end for
end for
```

6.6 H^2-conforming Hermite elements

Consider the Euler-Bernoulli beam problem for a simply supported beam reads

$$\begin{cases} \dfrac{d^2}{dx^2}\left[EI\dfrac{d^2w}{dx^2}\right] = q, \quad x \in (0,1) \\ w = 0, \quad x = 0 \\ w = 0, \quad x = 1 \\ EI\dfrac{d^2w}{dx^2} = 0, \quad x = 0 \\ EI\dfrac{d^2w}{dx^2} = 0, \quad x = 1 \end{cases} \tag{6.69}$$

The principle of virtual work gives

$$\begin{cases} w \in \mathcal{U} \\ b\left(w, v\right) = l\left(v\right), \quad \forall v \in \mathcal{U} \end{cases}, \tag{6.70}$$

where

$$b\left(w, v\right) = \int_0^1 \frac{d^2w}{dx^2} EI \frac{d^2v}{dx^2} dx, \tag{6.71}$$

$$l\left(v\right) = \int_0^1 qv dx, \quad \text{and} \tag{6.72}$$

$$\mathcal{U} = \left\{w \in H^2\left(0,1\right) \;:\; w(0) = w(1) = 0\right\}. \tag{6.73}$$

Next, we apply finite element discretization $w \approx w^h \in \mathcal{U}^h \subset \mathcal{U}$, i.e., dividing the domain into N subdomains $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, N-1$, where $0 = x_0 < x_1 < \ldots < x_N = 1$. We assign two degrees-of-freedom for each node, i.e., w^h and $\theta^h \equiv dw^h/dx$. Then, $w_{|K} \in X\left(K\right)$ reads

$$\begin{aligned} w_{|K}\left(\xi\right) = &\phi_0\left(\xi\right) w^h\left(x_i\right) + \phi_1\left(\xi\right)\theta^h\left(x_i\right) \\ &+ \phi_2\left(\xi\right) w^h\left(x_{i+1}\right) + \phi_3\left(\xi\right)\theta^h\left(x_{i+1}\right). \end{aligned} \tag{6.74}$$

Let $h = x_{i+1} - x_i$; shape functions ϕ_j are

$$\phi_0 = \frac{-\left(x - x_{i+1}\right)^2\left[-h + 2\left(x_i - x\right)\right]}{h^3}, \tag{6.75a}$$

$$\phi_1 = \frac{\left(x - x_i\right)\left(x - x_{i+1}\right)^2}{h^2}, \tag{6.75b}$$

$$\phi_2 = \frac{\left(x - x_i\right)^2\left[h + 2\left(x_{i+1} - x\right)\right]}{h^3}, \quad \text{and} \tag{6.75c}$$

$$\phi_3 = \frac{\left(x - x_i\right)^2\left(x - x_{i+1}\right)}{h^2}, \tag{6.75d}$$

which enforce C^1-continuity across element boundaries. The above shape functions belong to cubic Hermite shape functions. Namely, the element shape function space is the polynomial space of degree three: $X\left(K\right) = \mathcal{P}^3\left(K\right) = \text{span}\left\{1, x, x^2, x^3\right\}$.

TODO: shape functions plot

Example 6.6.1 (H^2-conformity) *Derive the continuity requirement for H^2-conforming finite element.*

The second-order derivative of a generalized function u reads

$$(u, \phi'') = (v, \phi)\,, \tag{6.76}$$

where v is the second-order derivative of u and $\phi \in C_0^\infty$ is a test function. In general, integration by parts gives

$$\int_0^1 u \frac{d^2\phi}{dx^2} dx = \int_0^1 \frac{d^2 u}{dx^2} \phi dx + \left[u \frac{d\phi}{dx} \right]_0^1 - \left[\frac{du}{dx} \phi \right]_0^1 . \tag{6.77}$$

Let $u_{|K} \in H^2(K)$ denote the function within each element. Then, we have

$$\begin{aligned}
\int_0^1 u \frac{d^2\phi}{dx^2} dx &= \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} \frac{d^2 u_{|K}}{dx^2} \phi dx + \sum_{i=0}^{N-1} \left[u_{|K} \frac{d\phi}{dx} \right]_{x_i}^{x_{i+1}} - \sum_{i=0}^{N-1} \left[\frac{du_{|K}}{dx} \phi \right]_{x_i}^{x_{i+1}} \\
&= \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} \frac{d^2 u_{|K}}{dx^2} \phi dx - \sum_{j=1}^{N-1} [u]_j \frac{d\phi}{dx} + \sum_{j=1}^{N-1} \left[\frac{du}{dx} \right]_j \phi \\
&= \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} \frac{d^2 u_{|K}}{dx^2} \phi dx + \int_0^1 \sum_{j=1}^{N-1} [u]_j \, \delta'(x - x_j)\, \phi dx \\
&\quad + \int_0^1 \sum_{j=1}^{N-1} \left[\frac{du}{dx} \right]_j \delta(x - x_j)\, \phi dx.
\end{aligned} \tag{6.78}$$

In the above, $[a]_i = a_{|K+1}(x_i) - a_{|K}(x_i)$ denotes the jump at an element interface $x = x_i$. Thus, the second-order derivative yields:

$$\frac{d^2 u}{dx^2} = \sum_{i=0}^{N-1} \frac{d^2 u_{|K}}{dx^2} + \sum_{j=1}^{N-1} [u]_j \, \delta'(x - x_j) + \sum_{j=1}^{N-1} \left[\frac{du}{dx} \right]_j \delta(x - x_j)\,. \tag{6.79}$$

Then, we can conclude that C^1-continuity is requried for an H^2-conforming element, i.e.,

$$[u]_j = \left[\frac{du}{dx} \right]_j = 0, \quad j = 1, 2, \ldots, N-1. \tag{6.80}$$

6.6.1 Nodal exactness

Next, we show that the approximation w^h is exact at the nodal points for the chosen set of shape functions. Consider a Green's function problem:

$$\begin{cases} EI\dfrac{d^4g}{dx^4} = \delta\left(x - x_o\right), \quad x \in (0,1) \\ g = 0, \quad x = 0 \\ g = 0, \quad x = 1 \\ EI\dfrac{d^2g}{dx^2} = 0, \quad x = 0 \\ EI\dfrac{d^2g}{dx^2} = 0, \quad x = 1 \end{cases}. \tag{6.81}$$

The solution is

$$\begin{aligned} EIg\left(x\right) =& \frac{1}{6}\left(x - x_o\right)^3 H\left(x - x_o\right) - \frac{1}{6}\left(1 - x_o\right)x^3 \\ &+ \frac{1}{6}\left(1 - x_o\right)x - \frac{1}{6}\left(1 - x_o\right)^3 x, \end{aligned} \tag{6.82}$$

where $H\left(x - x_o\right)$ is the Heaviside step function.

Notice that the Green's function is piecewise cubic. Therefore, $g \in \mathcal{U}^h$ when x_o is at one of nodes x_i. Then, the Galerkin orthogonality gives

$$\begin{aligned} 0 =& b\left(w - w^h, g\right) \\ =& \left(w - w^h, \delta\left(x - x_i\right)\right) \\ =& w\left(x_i\right) - w^h\left(x_i\right). \end{aligned} \tag{6.83}$$

Similarly, we can prove that the first-order derivative dw^h/dx is also exact at nodes. Here, the corresponding Green's function satisfies

$$\begin{cases} EI\dfrac{d^4g}{dx^4} = -\delta'\left(x - x_o\right), \quad x \in (0,1) \\ g = 0, \quad x = 0 \\ g = 0, \quad x = 1 \\ EI\dfrac{d^2g}{dx^2} = 0, \quad x = 0 \\ EI\dfrac{d^2g}{dx^2} = 0, \quad x = 1 \end{cases}, \tag{6.84}$$

where g is piecewise quadratic. Thus, $g \in \mathcal{U}^h$ when $x_o = x_i$, which gives

$$\begin{aligned} 0 =& b\left(w - w^h, g\right) \\ =& \left(w - w^h, -\delta'\left(x - x_i\right)\right) \\ =& \theta\left(x_i\right) - \theta^h\left(x_i\right). \end{aligned} \tag{6.85}$$

6.6.2 Accuracy of the higher-order derivatives

Let us take a further step by examining the accuracy of curvature, or the bending moment, i.e., EId^2w/dx^2. Because we expect a jump at an element interface, we can not apply the same approach above. Instead we find the optimal curvature points, or Barlow points, using Taylor expansions with remainders.

For simplicity, we work in the master coordinates $\xi \in [-1, 1]$:

$$\xi = \frac{2x - x_i - x_{i+1}}{h}. \tag{6.86}$$

The corresponding shape functions are

$$\begin{aligned}
\hat{\phi}_0(\xi) &= \frac{(\xi-1)^2(2+\xi)}{4}, && (6.87a)\\
\hat{\phi}_1(\xi) &= \frac{h(\xi+1)(\xi-1)^2}{8}, && (6.87b)\\
\hat{\phi}_2(\xi) &= \frac{(\xi+1)^2(2-\xi)}{4}, \quad \text{and} && (6.87c)\\
\hat{\phi}_3(\xi) &= \frac{h(\xi+1)^2(\xi-1)}{8}, && (6.87d)
\end{aligned}$$

where their second order derivatives are

$$\begin{aligned}
\hat{\phi}_0''(\xi) &= \frac{3}{2}\xi, && (6.88a)\\
\hat{\phi}_1''(\xi) &= \frac{h}{4}(3\xi-1), && (6.88b)\\
\hat{\phi}_2''(\xi) &= -\frac{3}{2}\xi, \quad \text{and} && (6.88c)\\
\hat{\phi}_3''(\xi) &= \frac{h}{4}(3\xi+1). && (6.88d)
\end{aligned}$$

The chain rule gives

$$\begin{aligned}
\frac{\partial}{\partial x} &= \frac{\partial}{\partial \xi}\frac{\partial \xi}{\partial x} = \frac{2}{h}\frac{\partial}{\partial \xi} \quad \text{and} && (6.89)\\
\frac{\partial^2}{\partial x^2} &= \left(\frac{2}{h}\right)^2 \frac{\partial^2}{\partial \xi^2}. && (6.90)
\end{aligned}$$

Then, the error of curvature can be expressed as

$$e''(x) = \left(\frac{2}{h}\right)^2 \hat{e}''(\xi) = \left(\frac{2}{h}\right)^2 \left[\hat{u}''(\xi) - \left(\hat{u}^h\right)''(\xi)\right], \tag{6.91}$$

where the nodal exactness gives

$$\begin{aligned}
\left(\hat{u}^h\right)''(\xi) =& \hat{\phi}_1''\hat{u}(-1) + \hat{\phi}_2''\frac{\partial \xi}{\partial x}\hat{u}'(-1) + \hat{\phi}_3''\hat{u}(1) + \hat{\phi}_4''\frac{\partial \xi}{\partial x}\hat{u}'(1)\\
=& \frac{3}{2}\xi\hat{u}(-1) + \frac{1}{2}(3\xi-1)\hat{u}'(-1) - \frac{3}{2}\xi\hat{u}(1) + \frac{1}{2}(3\xi+1)\hat{u}'(1). \qquad (6.92)
\end{aligned}$$

Next, perform Taylor expansions about $\xi = \alpha$ such that

$$\begin{aligned}\hat{u}(-1) =& \hat{u}(\alpha) + (-1-\alpha)\hat{u}'(\alpha) + \frac{1}{2}(-1-\alpha)^2\hat{u}''(\alpha) \\ &+ \frac{1}{6}(-1-\alpha)^3\hat{u}^{(3)}(\alpha) + \frac{1}{24}(-1-\alpha)^4\hat{u}^{(4)}(\alpha) \\ &+ \frac{1}{120}(-1-\alpha)^5\hat{u}^{(5)}(C_1), \qquad (6.93a)\\ \hat{u}'(-1) =& \hat{u}'(\alpha) + (-1-\alpha)\hat{u}''(\alpha) + \frac{1}{2}(-1-\alpha)^2\hat{u}^{(3)}(\alpha) \\ &+ \frac{1}{6}(-1-\alpha)^3\hat{u}^{(4)}(\alpha) + \frac{1}{24}(-1-\alpha)^4\hat{u}^{(5)}(C_2), \qquad (6.93b)\\ \hat{u}(1) =& \hat{u}(\alpha) + (1-\alpha)\hat{u}'(\alpha) + \frac{1}{2}(1-\alpha)^2\hat{u}''(\alpha) \\ &+ \frac{1}{6}(1-\alpha)^3\hat{u}^{(3)}(\alpha) + \frac{1}{24}(1-\alpha)^4\hat{u}^{(4)}(\alpha) \\ &+ \frac{1}{120}(1-\alpha)^5\hat{u}^{(5)}(C_3), \quad \text{and} \qquad (6.93c)\\ \hat{u}'(1) =& \hat{u}'(\alpha) + (1-\alpha)\hat{u}''(\alpha) + \frac{1}{2}(1-\alpha)^2\hat{u}^{(3)}(\alpha) \\ &+ \frac{1}{6}(1-\alpha)^3\hat{u}^{(4)}(\alpha) + \frac{1}{24}(1-\alpha)^4\hat{u}^{(5)}(C_4), \qquad (6.93d)\end{aligned}$$

where C_i are some values in the given domain. Plugging the above into the error formula we have

$$\hat{e}''(\alpha) = \frac{1}{6}\left(1-3\alpha^2\right)\hat{u}^{(4)}(\alpha) + \sum_i c_i u^{(5)}(C_i) \qquad (6.94)$$

or

$$e''(\bar{\alpha}) = \frac{h^2}{24}\left(1-3\alpha^2\right)u^{(4)}(\bar{\alpha}) + \mathcal{O}\left(h^3\right), \qquad (6.95)$$

where $\bar{\alpha} = x(\alpha)$. Thus, the optimal curvature points are located at

$$\xi = \pm\frac{1}{\sqrt{3}}, \qquad (6.96)$$

for which the corresponding error is of order $\mathcal{O}\left(h^3\right)$.

Note that these optimal curvature points correspond to the same integration points as those used in the two-point Gauss–Legendre quadrature. We also observe that the curvature is exact at the optimal points when each segment $[x_i, x_{i+1}]$ is uniformly loaded, i.e., q is constant. Additionally, the optimal point is everywhere when the segment is unloaded.

Exercise 6.6.1 (Accuracy of bar problem) *Repeat the preceding analyses on nodal exactness and derivative accuracy for the one-dimensional bar problem.*

6.7 Hierarchical elements

Hierarchical elements exhibit hierarchy in two aspects: polynomial order and dimension. The former means that higher-order shape functions include all lower-order ones, unlike in Lagrange elements. This property facilitates straightforward implementation of p-refinement. The latter refers to a dimensional hierarchy through traces, where higher-dimensional shape functions are constructed from lower-dimensional ones, providing a systematic framework for element definition.

6.7.1 Exact sequence elements

Exact sequence elements constitute a class of hierarchical elements in which the discrete spaces of element shape functions are consistent with the exact sequence of the underlying energy spaces. In addition to enabling a systematic construction of element shape functions, exact sequence elements ensure stability in mixed formulations. For a comprehensive discussion, see [Fuentes et al., 2015].

The four energy spaces form a sequence:

$$\mathbb{R} \xrightarrow{\text{id}} H^1 \xrightarrow{\text{grad}} H(\text{curl}) \xrightarrow{\text{curl}} H(\text{div}) \xrightarrow{\text{div}} L^2 \xrightarrow{0} \{0\}\,. \tag{6.97}$$

The above sequence is exact, i.e.,

$$\mathrm{R}(\text{grad}) = \mathrm{N}(\text{curl}) \quad \text{and} \quad \mathrm{R}(\text{curl}) = \mathrm{N}(\text{div})\,. \tag{6.98}$$

6.8 Summary on Lagrange shape functions

Table 6.1: Shape functions for two-noded line.

Geometry	
$\mu_0 = 1 - \xi$	$\operatorname{grad}\mu_0 = -1$
$\mu_1 = \xi$	$\operatorname{grad}\mu_1 = 1$
Shape Functions	
$\hat{\phi}_0 = \mu_0$	$\operatorname{grad}\hat{\phi}_0 = \operatorname{grad}\mu_0$
$\hat{\phi}_1 = \mu_1$	$\operatorname{grad}\hat{\phi}_1 = \operatorname{grad}\mu_1$

Table 6.2: Shape functions for three-noded line.

Geometry	
$\mu_0 = 2\left(\xi - \frac{1}{2}\right)(\xi - 1)$	$\operatorname{grad}\mu_0 = 2\left(2\xi - \frac{3}{2}\right)$
$\mu_1 = 2\xi\left(\xi - \frac{1}{2}\right)$	$\operatorname{grad}\mu_1 = 2\left(2\xi - \frac{1}{2}\right)$
$\mu_2 = 4\xi(1 - \xi)$	$\operatorname{grad}\mu_2 = 4(1 - 2\xi)$
Shape Functions	
$\hat{\phi}_0 = \mu_0$	$\operatorname{grad}\hat{\phi}_0 = \operatorname{grad}\mu_0$
$\hat{\phi}_1 = \mu_1$	$\operatorname{grad}\hat{\phi}_1 = \operatorname{grad}\mu_1$
$\hat{\phi}_2 = \mu_2$	$\operatorname{grad}\hat{\phi}_2 = \operatorname{grad}\mu_2$

Table 6.3: Shape functions for Q4 element.

Geometry	
$\mu_0^{\xi_1} = 1 - \xi_1$	$\operatorname{grad} \mu_0^{\xi_1} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$
$\mu_1^{\xi_1} = \xi_1$	$\operatorname{grad} \mu_1^{\xi_1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$
$\mu_0^{\xi_2} = 1 - \xi_2$	$\operatorname{grad} \mu_0^{\xi_2} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$
$\mu_1^{\xi_2} = \xi_2$	$\operatorname{grad} \mu_1^{\xi_2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
Shape Functions	
$\hat{\phi}_0 = \mu_0^{\xi_1} \mu_0^{\xi_2}$	$\operatorname{grad} \hat{\phi}_0 = \mu_0^{\xi_1} \operatorname{grad} \mu_0^{\xi_2} + \mu_0^{\xi_2} \operatorname{grad} \mu_0^{\xi_1}$
$\hat{\phi}_1 = \mu_1^{\xi_1} \mu_0^{\xi_2}$	$\operatorname{grad} \hat{\phi}_1 = \mu_1^{\xi_1} \operatorname{grad} \mu_0^{\xi_2} + \mu_0^{\xi_2} \operatorname{grad} \mu_1^{\xi_1}$
$\hat{\phi}_2 = \mu_1^{\xi_1} \mu_1^{\xi_2}$	$\operatorname{grad} \hat{\phi}_2 = \mu_1^{\xi_1} \operatorname{grad} \mu_1^{\xi_2} + \mu_1^{\xi_2} \operatorname{grad} \mu_1^{\xi_1}$
$\hat{\phi}_3 = \mu_0^{\xi_1} \mu_1^{\xi_2}$	$\operatorname{grad} \hat{\phi}_3 = \mu_0^{\xi_1} \operatorname{grad} \mu_1^{\xi_2} + \mu_1^{\xi_2} \operatorname{grad} \mu_0^{\xi_1}$

TODO: shape functions for 3D Lagrange elements.

Chapter 7

Time Integration

7.1 Separation of space and time variables

The finite element method (FEM) is generally applied to discretize the spatial domain and is not commonly used for temporal discretization. This is primarily because FEM is an implicit scheme that couples all degrees of freedom. Consequently, storing and factorizing the global system encompassing both spatial and temporal degrees of freedom would be computationally intractable and inconsistent with the causal nature of physical problems.

A common approach for a dynamic problem is to discretize space using FEM and discretize time using finite difference method (FDM). For example, consider a model one-dimensional wave equation:

$$\begin{cases} \dfrac{\partial}{\partial x}\left[EA\dfrac{\partial u(x,t)}{\partial x}\right] - \rho A\dfrac{\partial^2 u(x,t)}{\partial t^2} + f(x,t) = 0, & (x,t) \in \Omega \times I \\ u(x,t) = 0, & x = 0, L \\ u(x,t) = u_o(x), & t = 0 \\ \dfrac{\partial u(x,t)}{\partial t} = u_o'(x), & t = 0 \end{cases} . \tag{7.1}$$

The above equation describes wave propagation in a bar, when $\Omega = (0, L)$, $I = (0, T]$, E is Young's modulus, A is cross-sectional area, and ρ is mass density.

Then, the Galerkin method gives the following weak form

$$\begin{cases} u^h \in \mathcal{U}^h \\ b\left(u^h, v^h\right) = l\left(v^h\right), \quad \forall v^h \in \mathcal{U}^h \end{cases} , \tag{7.2}$$

where

$$b\left(u^h, v^h\right) = \int_0^L \frac{\partial u^h}{\partial x} EA \frac{\partial v^h}{\partial x} dx + \int_0^L \frac{\partial^2 u^h}{\partial t^2} \rho v^h dx \quad \text{and} \tag{7.3}$$

$$l\left(v^h\right) = \int_0^L f v^h dx. \tag{7.4}$$

The associated function space is

$$\mathcal{U}^h = \left\{ u^h(x,t) = a_i(t)\, g_i(x) \,:\, u^h\big|_t \in H^1(0,L),\, u^h(0,t) = u^h(L,t) = 0 \right\}. \tag{7.5}$$

Here, $g_i(x)$ is the finite element basis function. Its coefficient is denoted by $a_i(t)$, which is now a function in time.

The corresponding matrix equation reads

$$Ma'' + Ka = F, \tag{7.6}$$

where

$$\begin{aligned} M_{ij} &= \int_0^L g_i \rho A g_j dx, & \quad (7.7)\\ K_{ij} &= \int_0^L \frac{dg_i}{dx} EA \frac{dg_j}{dx} dx, \quad \text{and} & \quad (7.8)\\ F_i &= \int_0^L g_i f dx. & \quad (7.9) \end{aligned}$$

Thus. we have a linear system of ordinary differential equation in time. Thus, we are ready to use FDM for time integration.

Transforming a higher-order system into a first-order form is convenient, as it allows various numerical methods to be expressed in a unified and compact manner, and facilitates their straightforward adaptation to different problems. A representative first-order model problem is given by

$$\left\{ \begin{array}{ll} y' = f(y), & t \in I \\ y = y_o, & t = 0 \end{array} \right. . \tag{7.10}$$

If f is continuous with respect to t, then the solution to above reads

$$y(t) = y_o + \int_0^t f(y(\tau))\, d\tau. \tag{7.11}$$

For example, the above second-order system (7.6) can be written in the equivalent first-order form by defining

$$\begin{aligned} y &= \begin{pmatrix} u \\ u' \end{pmatrix}, & \quad (7.12)\\ y_o &= \begin{pmatrix} u_o \\ u'_o \end{pmatrix}, \quad \text{and} & \quad (7.13)\\ f(y) &= \begin{pmatrix} 0 \\ M^{-1}F \end{pmatrix} - \begin{bmatrix} 0 & I \\ M^{-1}K & 0 \end{bmatrix} \underbrace{\begin{pmatrix} u \\ u' \end{pmatrix}}_{=y}. & \quad (7.14) \end{aligned}$$

7.2 Finite difference methods

The basic principle of the finite difference method (FDM) is to approximate derivatives by divided differences, i.e.,

$$t \to \{t_n\},\ y(t_n) \sim y_n,\ n = 0, 1, \ldots,$$

where y_n denotes the discrete solution values to be determined. Table 7.1 summarizes several representative finite difference schemes.

Table 7.1: Representative finite difference schemes.

method	formula	remarks
(forward) Euler method	$\dfrac{y_{n+1}-y_n}{h} = f(y_n)$	explicit
backward Euler method	$\dfrac{y_{n+1}-y_n}{h} = f(y_{n+1})$	implicit
trapezoidal method	$\dfrac{y_{n+1}-y_n}{h} = \dfrac{1}{2}[f(y_{n+1})+f(y_n)]$	implicit
midpoint method	$\dfrac{y_{n+2}-y_n}{2h} = f(y_{n+1})$	explicit

A finite difference scheme is called a *one-step method* if $\forall n \geq 0$, u_{n+1} depends only on u_n. Here, the forward and backward Euler methods and the trapezoidal (or Crank-Nicolson) method are classified as one-step methods, while the midpoint method is a *multistep method.* Note that a multistep method requires a fictitious value u_{-1} to obtain u_1.

In terms of accuracy, the two Euler methods are first-order schemes with an error of $\mathcal{O}(h)$, whereas trapezoidal and midpoint methods are second-order schemes with an error of $\mathcal{O}(h^2)$. The order of accuracy is determined by the local truncation error (LTE). For $F(y_{n+k}, y_{n+k-1}, \ldots, y_n, h) = 0$ as a numerical method, define LTE_n as

$$\mathrm{LTE}_n = F(y_{n+k}, y_{n+k-1}, \ldots, y_n, h). \tag{7.15}$$

For example, LTE for the Euler method reads

$$\begin{aligned}\mathrm{LTE}_n =& F(y(t_{n+1}), y(t_n), h)\\ =& \frac{y(t_{n+1})-y(t_n)}{h} - f(y(t_n)) \text{ (using Taylor expansion and ODE)}\\ =& \frac{y(t_n)+hy'(t_n)+\frac{h^2}{2}y''(\tau_n)-y(t_n)}{h} - y'(t_n) \ \ (t_n \leq \tau_n \leq t_{n+1})\\ =& \frac{h}{2}y''(\tau_n) = \mathcal{O}(h).\end{aligned} \tag{7.16}$$

Implicit methods generally require solving a system of equations at each time step, while *explicit methods* compute u_{n+1} directly in terms of the previous values u_k, $k \leq n$. For many cases, implicit methods are more stable than explicit methods.

7.3 A-stability

A finite difference scheme is *absolutely stable* if for h fixed, u_n remains bounded as $n \to \infty$. Alternatively, the absolute stability is defined on the following test problem [Quarteroni et al., 2006]:

$$\begin{cases} y'(t) = \lambda y(t), & t > 0 \\ y(0) = 1 \end{cases}, \tag{7.17}$$

where $\lambda \in \mathbb{C}$. Then, a finite difference scheme for approximating the test problem (7.17) is absolutely stable if

$$|y_n| \to 0 \text{ as } n \to \infty. \tag{7.18}$$

Note that the solution to the test problem is $y = e^{\lambda t}$.

The absolute stability is generally depends on h and λ, where the *region of absolute stability* is defined as

$$\mathcal{A} = \{z = h\lambda \in \mathbb{C} \ : \ |y_n| \to 0 \text{ as } n \to \infty\}. \tag{7.19}$$

For example, the forwards Euler method for approximating the test problem gives $y_{n+1} = y_n + h\lambda y_n$ or

$$y_n = (1 + \lambda h)^n, \quad n \geq 0. \tag{7.20}$$

The absolute stability holds iff

$$|1 + h\lambda| < 1, \tag{7.21}$$

i.e., when λh lies inside the unit circle centered at $(-1, 0)$ in the complex plane. On the other hand, the backward Euler method gives

$$y_n = (1 - \lambda h)^{-n}, \quad n \geq 0. \tag{7.22}$$

Thus, the absolute stability holds iff

$$|1 - h\lambda| > 1. \tag{7.23}$$

The regions of absolute stability for the above two methods are shown in Figure 7.1.

In addition, a method is called *A-stable* if

$$\mathcal{A} \cap \mathbb{C}_- = \mathbb{C}_-, \tag{7.24}$$

where

$$\mathbb{C}_- = \{z \in \mathbb{C} \ : \ \mathrm{Re}\,(z) < 0\}. \tag{7.25}$$

A-stability is also called unconditional absolute stability. The backward Euler and trapezoidal methods are A-stable. Note that there is no explicit method that is A-stable.

Exercise 7.3.1 (Stability of Euler method) *Approximate the solution of*

$$\begin{cases} y'(t) = -5y(t), & t > 0 \\ y(0) = 1 \end{cases}. \tag{7.26}$$

using both the forward Euler and backward Euler methods. Present the numerical results for various sizes of h, and discuss the observations with respect to each method's region of absolute stability.

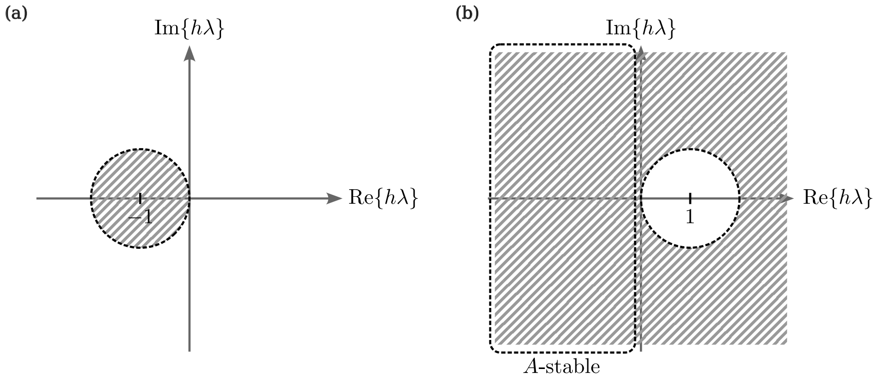

Figure 7.1: Region of absolute stability. (a) forward Euler method. (b) backward Euler method showing A-stability.

7.4 Runge-Kutta methods

The Runge–Kutta (RK) methods are high-order multi-stage time-integration schemes. Their general form is expressed as

$$y_{n+1} = y_n + h\sum_{k=1}^{S} b_k K_k, \quad \text{where} \tag{7.27}$$

$$K_k = f\left(y_n + h\sum_{j=1}^{S} a_{jk} K_j\right). \tag{7.28}$$

The method is explicit when $a_{jk} = 0$ for all $j \geq k$. Here, S denotes the number of stages, i.e., the number of function evaluations performed per time step. Accordingly, the Runge–Kutta methods are classified as multi-stage methods.

For example, a second-order Runge–Kutta method can be written as

$$y_{n+1} = y_n + h\left(\frac{1}{2}K_1 + \frac{1}{2}K_2\right), \tag{7.29}$$

$$K_1 = f\left(y_n + 0\right), \tag{7.30}$$

$$K_2 = f\left(y_n + hK_1\right). \tag{7.31}$$

This scheme can be derived from the trapezoidal rule combined with the Euler method, namely,

$$\begin{aligned}\frac{y_{n+1} - y_n}{h} &= \frac{1}{2}\left[f\left(y_n\right) + f\left(y_{n+1}\right)\right] \quad (\text{use Euler to determine } f\left(y_{n+1}\right)) \\ &= \frac{1}{2}\left[f\left(y_n\right) + f\left(y_n + hf\left(y_n\right)\right)\right]. \end{aligned} \tag{7.32}$$

The standard 4th-order explicit RK method is

$$y_{n+1} = y_n + h\left[\frac{1}{6}K_1 + \frac{2}{6}K_2 + \frac{2}{6}K_3 + \frac{1}{6}K_4\right], \tag{7.33}$$

where

$$\begin{aligned}
K_1 &= f\left(y_n\right), && (7.34a)\\
K_2 &= f\left(y_n + \frac{h}{2}K_1\right), && (7.34b)\\
K_3 &= f\left(y_n + \frac{h}{2}K_2\right), && (7.34c)\\
K_4 &= f\left(y_n + hK_3\right). && (7.34d)
\end{aligned}$$

In general, the coefficients b_k and a_{jk} are determined through a Taylor series expansion of the numerical scheme and the ODE. The coefficients are selected so that the leading terms of the local truncation error (LTE) are canceled. The single-stage scheme corresponds to the Euler method. In practice, the number of stages S is typically chosen between 1 and 6, since eliminating higher-order LTE terms becomes increasingly difficult for $S > 6$.

Example 7.4.1 (4th-order RK for 2nd-order problems) *Derive the standard 4th-order RK method for the nonlinear second-order system:*

$$Mu'' + Cu' + g\left(u\right) = F, \tag{7.35}$$

where M is invertible.

Convering the above into first-order system $y' = f\left(y\right)$, we have

$$y = \begin{pmatrix} u \\ v \end{pmatrix} \quad \text{and} \quad f\left(y\right) = \begin{pmatrix} v \\ M^{-1}\left[F - Cv - g\left(u\right)\right] \end{pmatrix}. \tag{7.36}$$

Let $K_i = \left(K_i^u,\ K_i^v\right)^T$, (7.34) is expanded by

$$\begin{aligned}
\begin{pmatrix} K_1^u \\ K_1^v \end{pmatrix} &= \begin{pmatrix} v_n \\ M^{-1}\left[F - Cv_n - g\left(u_n\right)\right] \end{pmatrix}, && (7.37a)\\
\begin{pmatrix} K_2^u \\ K_2^v \end{pmatrix} &= \begin{pmatrix} v_n + \frac{h}{2}K_1^v \\ M^{-1}\left[F - C\left(v_n + \frac{h}{2}K_1^v\right) - g\left(u_n + \frac{h}{2}K_1^u\right)\right] \end{pmatrix}, && (7.37b)\\
\begin{pmatrix} K_3^u \\ K_3^v \end{pmatrix} &= \begin{pmatrix} v_n + \frac{h}{2}K_2^v \\ M^{-1}\left[F - C\left(v_n + \frac{h}{2}K_2^v\right) - g\left(u_n + \frac{h}{2}K_2^u\right)\right] \end{pmatrix}, && (7.37c)\\
\begin{pmatrix} K_4^u \\ K_4^v \end{pmatrix} &= \begin{pmatrix} v_n + hK_3^v \\ M^{-1}\left[F - C\left(v_n + hK_3^v\right) - g\left(u_n + hK_3^u\right)\right] \end{pmatrix}. && (7.37d)
\end{aligned}$$

Then, we have

$$\begin{aligned}
u_{n+1} &= u_n + h\left[\frac{1}{6}K_1^u + \frac{2}{6}K_2^u + \frac{2}{6}K_3^u + \frac{1}{6}K_4^u\right] \quad \text{and} && (7.38)\\
v_{n+1} &= v_n + h\left[\frac{1}{6}K_1^v + \frac{2}{6}K_2^v + \frac{2}{6}K_3^v + \frac{1}{6}K_4^v\right]. && (7.39)
\end{aligned}$$

7.5 Newmark method

Newmark method on

$$Mu'' + Cu' + g(u) = F \tag{7.40}$$

reads

$$u_{n+1} = u_n + hv_n + \frac{h^2}{2}\left[(1-2\beta)\,a_n + 2\beta a_{n+1}\right] \quad \text{and} \tag{7.41}$$

$$v_{n+1} = v_n + (1-\gamma)\,ha_n + \gamma h a_{n+1}, \tag{7.42}$$

where

$$a_n = M^{-1}\left[F - Cv_n - g(u_n)\right]. \tag{7.43}$$

In the above, β and γ are parameters. Typical choices gives

- average acceleration (trapezoidal method): $\gamma = 1/2$ and $\beta = 1/4$
- linear acceleration: $\gamma = 1/2$ and $\beta = 1/6$
- explicit central difference: $\gamma = 1/2$ and $\beta = 0$

Bibliography

[Ciarlet, 2002] Ciarlet, P. (2002). *The Finite Element Method for Elliptic Problems.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

[Cook, 2001] Cook, R. (2001). *Concepts and Applications of Finite Element Analysis.* Wiley.

[Demkowicz, 2023] Demkowicz, L. (2023). *Mathematical Theory of Finite Elements.* Computational science and engineering. Society for Industrial and Applied Mathematics.

[Engquist, 2014] Engquist, B. (2014). Lecture notes of numerical analysis: Differential equations. The University of Texas at Austin.

[Fuentes et al., 2015] Fuentes, F., Keith, B., Demkowicz, L., and Nagaraj, S. (2015). Orientation embedded high order shape functions for the exact sequence elements of all shapes. *Computers & Mathematics with applications*, 70(4):353–458.

[Hughes, 2012] Hughes, T. (2012). *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis.* Dover Civil and Mechanical Engineering. Dover Publications.

[Lee, 2022] Lee, H. S. (2022). Introduction to finite element method. Seoul National University.

[Oden and Demkowicz, 2017] Oden, J. T. and Demkowicz, L. (2017). *Applied functional analysis.* Chapman and Hall/CRC.

[Quarteroni et al., 2006] Quarteroni, A., Sacco, R., and Saleri, F. (2006). *Numerical Mathematics.* Texts in Applied Mathematics. Springer Berlin Heidelberg.