AER1418: Variational Methods for PDEs

Lecture Notes

Version 1.0 (Winter 2022)

Masayuki Yano

University of Toronto
Institute for Aerospace Studies

Acknowledgment

I have prepared this lecture notes for a graduate course AER1418 Variational Methods for PDEs taught at the University of Toronto. I would like to thank Prof. Anthony Patera of MIT for many fruitful discussions during the initial development of this course and for general discussions on pedagogy over the years. I would also like to thank all of my past AER1418 students who have provide many helpful feedback and improved the notes.

Contents

Lecture 1

Introduction: Poisson's equation in one dimension

1.1 Introduction

In this lecture, we provide a brief overview of the variational formulation and the associated finite element approximation of partial differential equations using a concrete example: one-dimensional Poisson's equation. The goal is to provide an accessible overview that illustrates the main ideas without complexities associated with higher dimensions and more general equations; we also defer some of the technical discussions to later lectures.

1.2 Model problem: strong form

We consider a taut string with fixed ends subjected to a distributed transverse load as shown in Figure 1.1. Given an appropriate normalization, the shape of the string can be modeled as the solution to a one-dimensional Poisson's equation on $\Omega \equiv (0,1)$:

$$\begin{aligned} -\frac{d^2u}{dx^2} &= f \qquad \text{in } \Omega, \\ u(x=0) &= 0, \\ u(x=1) &= 0, \end{aligned} \tag{1.1}$$

where f is associated with the transverse load. For an integrable f, we may integrate the ODE twice to confirm the existence and uniqueness of the solution. We refer to this particular form of the problem as the *strong form*. The name derives from the fact the equation is enforced strongly in the point-wise sense for each x in Ω, which can be contrasted with the *weak form* introduced in the next section.

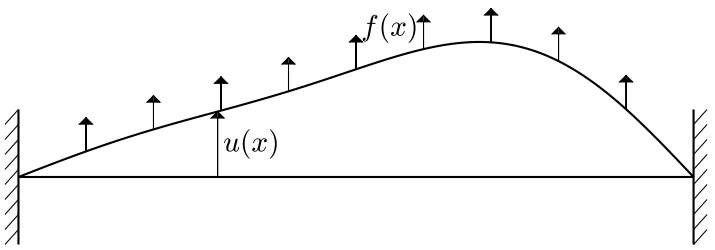


Figure 1.1: Taut-string problem modeled by Poisson's equation.

1.3 Variational formulation (or weak formulation)

We now consider a different form of Poisson's equation (1.1) that is (i) more general than the strong form and (ii) amenable to finite element discretization. To obtain this new form of (1.1), we use the *weighted residual method*. To begin, we choose a (sufficiently regular) test function such that $v(x=0) = v(x=1) = 0$, multiply (1.1) by this function, and integrate the expression to obtain

$$\int_\Omega v\left(-\frac{d^2u}{dx^2}\right)dx = \int_\Omega vf dx. \tag{1.2}$$

We then apply integration by parts to the left hand side to obtain

$$\int_\Omega \frac{dv}{dx}\frac{du}{dx}dx - \cancel{\left[v\frac{du}{dx}\right]_{x=0}^{1}} = \int_\Omega vf dx; \tag{1.3}$$

note that the boundary term vanishes because we require $v(x=0) = v(x=1) = 0$. If u is the solution to (1.1), then we expect the expression (1.2), and in turn (1.3), to hold for *any* test function v. In fact, the *variational problem* seeks the solution u which satisfies (1.3) for all suitable test functions v. We may speculate that, given a suitably large set of test functions, the solution u is unique and coincides with the solution to the strong form of Poisson's equation (1.1).

We now formalize the above procedure. To this end, we first introduce a linear space

$$\mathcal{V} \equiv \{v \mid v \text{ is continuous, } \int_\Omega \left(\frac{dv}{dx}\right)^2 dx \text{ is bounded, and } v(x=0) = v(x=1) = 0\}; \tag{1.4}$$

here, first two conditions are related to the regularity of the solution sought, and the last condition imposes the boundary conditions. We note that $v \in \mathcal{V}$ need not be twice continuously differentiable. In fact $\frac{dv}{dx}$ does not have to be even bounded; it just needs to be square integrable.

We next introduce a *linear form* $\ell : \mathcal{V} \to \mathbb{R}$ such that

$$\ell(v) \equiv \int_\Omega vf dx \quad \forall v \in \mathcal{V}. \tag{1.5}$$

The form $\ell : \mathcal{V} \to \mathbb{R}$ is called a *linear form* because it is linear in its argument in the sense that

(i) $\ell(\alpha w) = \alpha\ell(w) \quad \forall w \in \mathcal{V},\ \forall \alpha \in \mathbb{R}$;

(ii) $\ell(w+v) = \ell(w) + \ell(v) \quad \forall w, v \in \mathcal{V}$.

In other words,

$$\ell(\alpha w + \beta v) = \alpha \ell(w) + \beta \ell(v) \quad \forall w, v \in \mathcal{V},\ \forall \alpha, \beta \in \mathbb{R}.$$

We also define a *bilinear form* $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ such that

$$a(w, v) \equiv \int_\Omega \frac{dv}{dx}\frac{dw}{dx} dx \quad \forall w, v \in \mathcal{V}. \tag{1.6}$$

The form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is called a *bilinear form* because

(i) for any fixed $\tilde{v}$, $a(w, \tilde{v})$ is a linear form in w;

(ii) for any fixed $\tilde{w}$, $a(\tilde{w}, v)$ is a linear form in v.

Given the linear and bilinear forms, we can concisely state our *variational formulation* of the problem as follows: find $u \in \mathcal{V}$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathcal{V}. \tag{1.7}$$

This variational formulation of the problem is also called the *weak formulation*. The space $\mathcal{V}$ to which the solution u belongs is called the *trial space*; the space $\mathcal{V}$ to which the test function v belongs is called the *test space*. (While we choose the same trial and test spaces in this example, the two spaces need not be the same in general.) A variational form which uses the same function space for the trial and test spaces is called the *Galerkin formulation*; our variational form (1.7) is a Galerkin formulation because both the trial and test spaces are $\mathcal{V}$ defined by (1.4). The variational problem has a unique solution; we will study the well-posedness of the problem in subsequent lectures.

We can readily show that the solution to the strong from (1.1) solves the variational form (1.7). To see this, we integrate by parts and observe that, for all $v \in \mathcal{V}$,

$$a(u, v) - \ell(v) \equiv \int_\Omega \frac{dv}{dx}\frac{du}{dx} dx - \int_\Omega v f dx = \int_\Omega v \underbrace{\left(-\frac{d^2u}{dx^2} - f\right)}_{=0 \text{ as } u \text{ solves (1.1)}} dx + \underbrace{\left[v\frac{du}{dx}\right]_{x=0}^{1}}_{=0 \text{ by BC for } v} = 0.$$

Hence, the solution to the strong form (1.1) satisfies the weak form (1.7). However, the converse is not necessarily true. In fact, the variational form admits more general loads f and associated solutions than the strong form. For instance, a Dirac delta function, which corresponds to a point load, is an admissible load for the variational formulation (1.7) and there exists a unique solution to the problem; however, the strong from (1.1) is not well defined for a Dirac delta function f because f does not have well-defined point-wise values.

1.4 Minimization formulation

We now introduce a *minimization form*, which is closely related to the variational form (1.7). We note that not all boundary value problems possess a minimization form; only those problems with an intrinsic energy, such as our model problem (1.1), possess a minimization form. To obtain a minimization form, we introduce a functional $J : \mathcal{V} \to \mathbb{R}$ given by

$$J(w) \equiv \frac{1}{2}\int_\Omega \left(\frac{dw}{dx}\right)^2 dx - \int_\Omega f w dx \quad \forall w \in \mathcal{V}, \tag{1.8}$$

where $\mathcal{V}$ is defined by (1.4). Our *minimization formulation* is as follows: find $u \in \mathcal{V}$ such that

$$u = \arg\min_{w \in \mathcal{V}} J(w). \tag{1.9}$$

For a physical system with intrinsic energy, such as the taut-string problem in Section 1.1, the functional J represents the total energy in the system; $\int_\Omega (\frac{\partial w}{\partial x})^2 dx$ is the internal energy and $\int_\Omega f w dx$ is the external work. Our minimization statement is hence a statement of energy minimization at the equilibrium.

We can readily show that $u \in \mathcal{V}$ is the solution to the variational problem (1.7) if and only if it is the solution to the minimization problem (1.9). We first show that if $u \in \mathcal{V}$ solves the variational problem (1.7), then it satisfies the minimization problem (1.9). Suppose $u \in \mathcal{V}$ is the solution to the variational problem (1.7). Then, for $w = u + v$ for any $v \in \mathcal{V}$, we obtain

$$J(w) = J(u+v) = J(u) + \underbrace{\int_\Omega \frac{dv}{dx}\frac{du}{dx} dx - \int_\Omega f v dx}_{\text{(I)}} + \underbrace{\frac{1}{2}\int_\Omega \left(\frac{dv}{dx}\right)^2 dx}_{\text{(II)}}.$$

The term (I) vanishes because by (1.7)

$$\int_\Omega \frac{dv}{dx}\frac{du}{dx} dx - \int_\Omega f v dx = a(u,v) - \ell(v) = 0 \quad \forall v \in \mathcal{V}.$$

We now analyze (II). We first note that the term is greater than 0 for all non-constant function (i.e. $\frac{dv}{dx} \neq 0$); (II) is zero only if v is a constant function. We next note that for a constant function to satisfy the boundary conditions of $\mathcal{V}$, $v(x=0) = v(x=1) = 0$, the function must be the zero function: $v = 0$. Hence, we conclude that (II) is greater than 0 for any non-zero function. It thus follows that

$$J(w) = J(u+v) > J(u) \quad \forall v \neq 0;$$

hence the solution $u \in \mathcal{V}$ of the variational problem (1.7) solves the minimization problem (1.9).

We now show the converse: if $u \in \mathcal{V}$ solves the minimization problem (1.9), then it satisfies the variational problem (1.7). Suppose $u \in \mathcal{V}$ is the solution to the minimization problem (1.9). We know that the minimizer $u \in \mathcal{V}$ must satisfy the stationarity condition

$$J'(u;v) \equiv \lim_{\epsilon \to 0} \frac{1}{\epsilon}(J(u+\epsilon v) - J(u)) = 0 \quad \forall v \in \mathcal{V};$$

i.e., the Fréchet derivative (i.e., directional derivative) about u in any direction v should be 0. The stationarity condition can be restated as

$$\begin{aligned} J'(u;v) &\equiv \lim_{\epsilon \to 0} \frac{1}{\epsilon}(J(u+\epsilon v) - J(u)) \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon}\left(\int_\Omega \epsilon \frac{dv}{dx}\frac{du}{dx} dx - \int_\Omega \epsilon f v dx + \frac{1}{2}\int_\Omega \epsilon^2 \left(\frac{dv}{dx}\right)^2 dx\right) \\ &= \int_\Omega \frac{dv}{dx}\frac{du}{dx} dx - \int_\Omega f v dx = a(u,v) - \ell(v) = 0 \quad \forall v \in \mathcal{V}, \end{aligned}$$

which is exactly our variational problem (1.7). Hence we conclude that $u \in \mathcal{V}$ is the solution to the variational problem (1.7) if and only if it is the solution to the minimization problem (1.9); the two

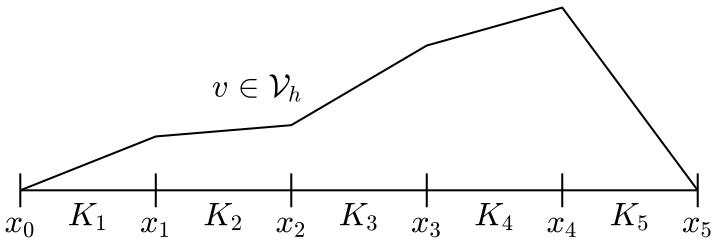


Figure 1.2: Linear finite element space for $N = 4$.

problems are equivalent. It follows that, as noted in Section 1.3, if the solution u solves the strong form (1.1), then it also the solution to the minimization form (1.9); however, the minimization form admits more general loads f and the associated solutions than the strong form. We will soon see that we can derive a finite element approximation from either (1.7) or (1.9).

1.5 Finite element approximation

In order to construct a finite element (FE) approximation, we must first choose a suitable subspace of $\mathcal{V}$ that well approximates $\mathcal{V}$ and is amenable to computer implementation. To this end, we first *triangulate* the domain $\Omega \equiv (0, 1)$ into $N + 1$ non-overlapping segments; we introduce points

$$0 \equiv x_0 < x_1 < \cdots < x_N < x_{N+1} \equiv 1$$

and segments

$$K_i \equiv (x_{i-1}, x_i) \quad i = 1, \ldots, N + 1.$$

We denote the *triangulation* of the domain Ω, which comprises collection of line segments, by

$$\mathcal{T}_h \equiv \{K_i\}_{i=1}^{N+1}. \tag{1.10}$$

We denote the length of each segment by $h_i \equiv x_i - x_{i-1}$, $i = 1, \ldots, N + 1$; a triangulation $\mathcal{T}_h$ is characterized by the maximum segment length $h \equiv \max_{i=\{1,\ldots,N+1\}} h_i$.

We now introduce a space of piecewise linear functions defined over our triangulation $\mathcal{T}_h$ such that they also belong to $\mathcal{V}$:

$$\mathcal{V}_h \equiv \{v \in \mathcal{V} \mid v|_{K_i} \in \mathbb{P}^1(K_i),\ i = 1, \ldots, N + 1\}. \tag{1.11}$$

We make a few observations about this particular space. First, we require that v is in $\mathcal{V}$ defined by (1.4): (i) v must be continuous, (ii) $\frac{dv}{dx}$ must be piecewise continuous and bounded, and (iii) v must vanish at the boundaries. Second, we require v restricted to segment K_i to be in $\mathbb{P}^1(K_i)$; here, $\mathbb{P}^1(K_i)$ denotes the space of linear polynomials over K_i. Hence, any function in $\mathcal{V}_h \subset \mathcal{V}$ is continuous, piecewise linear, and vanishes at the boundaries. An example of a function in $\mathcal{V}_h$ is shown in Figure 1.2.

We can now state our FE approximation problem in either the variational form or the minimization form. The FE variational formulation of our original problem (1.7) is as follows: find $u_h \in \mathcal{V}_h$ such that

$$a(u_h, v) = \ell(v) \quad \forall v \in \mathcal{V}_h, \tag{1.12}$$

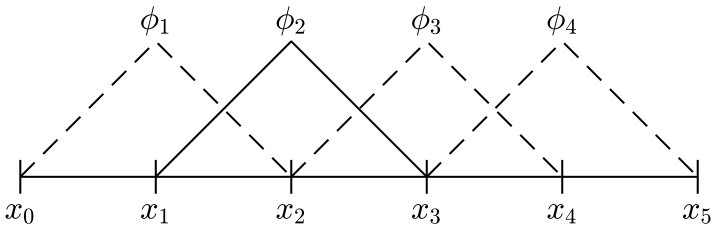


Figure 1.3: Basis functions for a linear finite element space for $N = 4$.

where $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and $\ell : \mathcal{V} \to \mathbb{R}$ are the bilinear form (1.6) and linear form (1.5), respectively. In words, we seek the solution in and test against the finite-dimensional (hence computable) subspace $\mathcal{V}_h$ of $\mathcal{V}$.

Similarly, the FE minimization form is as follows: find $u_h \in \mathcal{V}_h$ such that

$$u_h = \arg\min_{v \in \mathcal{V}_h} J(v), \tag{1.13}$$

where $J : \mathcal{V} \to \mathbb{R}$ is the energy functional (1.8). In words, we seek the minimizer of J in the subspace $\mathcal{V}_h$ of $\mathcal{V}$. As before, $u_h \in \mathcal{V}_h$ is the solution to the FE minimization problem (1.13) if and only if it is the solution to the FE variational problem (1.12). We will henceforth refer to u_h as the *finite element solution*.

1.6 Finite element approximation: implementation

We now wish to develop a computationally tractable formulation such that we can implement (1.12) (or equivalently (1.13)) and find an approximation to the original problem (1.7) (or equivalently (1.9)). To this end, we introduce a *basis* for $\mathcal{V}_h$ defined in (1.11). A convenient basis for $\mathcal{V}_h$ is a *Lagrange basis* $\{\phi_i\}_{i=1}^N$ comprised piecewise linear polynomials such that

$$\phi_i(x_j) = \delta_{ij} \equiv \begin{cases} 1 & j = i \\ 0 & j \neq i \end{cases}, \quad j = 0, \ldots, N+1;$$

in words, ϕ_i is a continuous piecewise linear function that is takes the value of 1 at the interpolation point x_i and the value of 0 at all other interpolation points. These basis functions are shown in Figure 1.3.

We can show that $\{\phi_i\}_{i=1}^N$ is indeed a basis for $\mathcal{V}_h$: the set (i) is linearly independent and (ii) spans $\mathcal{V}_h$. To show the set is linearly independent, we must show that $\sum_{i=1}^n c_i \phi_i(x) = 0$ implies $c_i = 0$, $i = 1, \ldots, N$; the statement holds because, for $\sum_{i=1}^n c_i \phi_i(x) = 0$, we must have $\sum_{i=1}^n c_i \phi_i(x_j) = \sum_{i=1}^n c_i \delta_{ij} = c_j = 0$, $j = 1, \ldots, N$. To show the set spans $\mathcal{V}_h$, we observe that any $v \in \mathcal{V}_h$ can be expressed as

$$v = \sum_{i=1}^N \hat{v}_i \phi_i$$

for $\hat{v}_i \equiv v(x_i)$. (The coefficients are also unique because the set is linearly independent.) We hence conclude that $\{\phi_i\}_{i=1}^N$ is a basis for $\mathcal{V}_h$.

We now restate the FE variational problem (1.12) using the basis and associated coefficients. Specifically, we represent $u_h \in \mathcal{V}_h$ and $v \in \mathcal{V}_h$ as

$$u_h \equiv \sum_{j=1}^{N} \hat{u}_{h,j}\phi_j \tag{1.14}$$
$$v \equiv \sum_{i=1}^{N} \hat{v}_i\phi_i$$

for some $\hat{u}_h \in \mathbb{R}^N$ and $\hat{v} \in \mathbb{R}^N$, and consider the following equivalent problem: find $\hat{u}_h \in \mathbb{R}^N$ such that

$$a(\sum_{j=1}^{N} \hat{u}_{h,j}\phi_j, \sum_{i=1}^{N} \hat{v}_i\phi_i) - \ell(\sum_{i=1}^{N} \hat{v}_i\phi_i) = \sum_{i=1}^{N}\sum_{j=1}^{N} \hat{v}_i a(\phi_j,\phi_i)\hat{u}_{h,j} - \sum_{i=1}^{N} \hat{v}_i\ell(\phi_i) = 0 \quad \forall \hat{v} \in \mathbb{R}^N. \tag{1.15}$$

Here, we have appealed to the bilinearity and linearity of $a(\cdot,\cdot)$ and $\ell(\cdot)$, respectively. We then note that in (1.15) we can replace the condition $\forall \hat{v} \in \mathbb{R}^N$ with an equivalent condition that the statement holds for all $\hat{v} \in \{e_i\}_{i=1}^N$, where $\{e_i\}_{i=1}^N$ is the canonical basis of $\mathbb{R}^N$ (i.e., $e_i \in \mathbb{R}^N$ has 1 in the i-th entry and 0 elesewhere). Then, we can restate (1.15) as follows: find $\hat{u}_h \in \mathbb{R}^N$ such that

$$\sum_{j=1}^{N} a(\phi_j,\phi_i)\hat{u}_{h,j} = \ell(\phi_i) \quad \forall i = 1,\dots,N. \tag{1.16}$$

We can also rewrite the linear system in the matrix form:

$$\underbrace{\begin{pmatrix} a(\phi_1,\phi_1) & \cdots & a(\phi_N,\phi_1) \\ \vdots & \ddots & \vdots \\ a(\phi_1,\phi_N) & \cdots & a(\phi_N,\phi_N) \end{pmatrix}}_{\hat{A}_h \in \mathbb{R}^{N\times N}} \underbrace{\begin{pmatrix} \hat{u}_{h,1} \\ \vdots \\ \hat{u}_{h,N} \end{pmatrix}}_{\hat{u}_h \in \mathbb{R}^N} = \underbrace{\begin{pmatrix} \ell(\phi_1) \\ \vdots \\ \ell(\phi_N) \end{pmatrix}}_{\hat{f}_h \in \mathbb{R}^N},$$

or, more concisely,

$$\hat{A}_h\hat{u}_h = f_h.$$

The matrix $\hat{A}_h$ is called the *stiffness matrix* and the vector f_h is called the *load vector*.

We now take a closer look at the matrix $\hat{A}_h \in \mathbb{R}^{N\times N}$ associated with our particular choice of the basis $\{\phi_i\}_{i=1}^N$. We consider four distinct parts of the matrix: the main diagonal, superdiagonal, subdiagonal, and all other entries. The diagonal entries are given by

$$a(\phi_i,\phi_i) = \int_{x_{i-1}}^{x_{i+1}} \frac{d\phi_i}{dx}\frac{d\phi_i}{dx}dx = \int_{x_{i-1}}^{x_i} \left(\frac{1}{h_i}\right)^2 dx + \int_{x_i}^{x_{i+1}} \left(-\frac{1}{h_{i+1}}\right)^2 dx = \frac{1}{h_i} + \frac{1}{h_{i+1}}, \quad i = 1,\dots,N.$$

The superdiagonal entries are given by

$$a(\phi_{i+1},\phi_i) = \int_{x_i}^{x_{i+1}} \frac{d\phi_{i+1}}{dx}\frac{d\phi_i}{dx}dx = \int_{x_i}^{x_{i+1}} \left(\frac{1}{h_{i+1}}\right)\left(-\frac{1}{h_{i+1}}\right) dx = -\frac{1}{h_{i+1}}, \quad i = 1,\dots,N-1.$$

The subdiagonal entries are given by

$$a(\phi_i,\phi_{i+1}) = \int_{x_i}^{x_{i+1}} \frac{d\phi_i}{dx}\frac{d\phi_{i+1}}{dx}dx = \int_{x_i}^{x_{i+1}} \left(-\frac{1}{h_{i+1}}\right)\left(\frac{1}{h_{i+1}}\right) dx = -\frac{1}{h_{i+1}}, \quad i = 1,\dots,N-1.$$

All other entries are zero because ϕ_i and ϕ_j do not overlap for $|i-j|>1$. The stiffness matrix is hence given by

$$\hat{A}_h = \begin{pmatrix} h_1^{-1}+h_2^{-1} & -h_2^{-1} & & & & \\ -h_2^{-1} & h_2^{-1}+h_3^{-1} & -h_3^{-1} & & & \\ & \ddots & \ddots & \ddots & & \\ & & -h_{N-1}^{-1} & h_{N-1}^{-1}+h_N^{-1} & -h_N^{-1} \\ & & & -h_N^{-1} & h_N^{-1}+h_{N+1}^{-1} \end{pmatrix}.$$

We observe that the matrix is *sparse* and in particular *tridiagonal*. Moreover, this matrix is symmetric positive definite. The symmetry is obvious from inspection; the positive definiteness follows from

$$\begin{aligned} \hat{v}^T \hat{A}_h \hat{v} &= h_1^{-1}\hat{v}_1^2 + \sum_{i=1}^{N-1} h_{i+1}^{-1}(\hat{v}_i^2 - 2\hat{v}_i\hat{v}_{i+1} + \hat{v}_{i+1}^2) + h_{N+1}^{-1}\hat{v}_N^2 \\ &= h_1^{-1}\hat{v}_1^2 + \sum_{i=1}^{N-1} h_{i+1}^{-1}(\hat{v}_i^2 - \hat{v}_{i+1})^2 + h_{N+1}^{-1}\hat{v}_N^2 > 0 \quad \forall \hat{v} \neq 0. \end{aligned}$$

Hence the solution exists and is unique. The storage requirement for the tridiagonal system is $\mathcal{O}(N)$, and the solution to $\hat{A}_h\hat{u}_h = f_h$ can be obtained using the Thomas algorithm, which is a form of Gaussian elimination, in $\mathcal{O}(N)$ floating point operations. Given the solution $\hat{u}_h \in \mathbb{R}^N$ to the linear system, our finite element solution $u_h \in \mathcal{V}_h$ to the FE variational formulation (1.12) (or equivalently (1.13)) is given by the representation (1.14).

If the nodes are equispaced so that $h_i = h$, $\forall i = 1, \ldots, N$, the stiffness matrix simplifies to

$$\hat{A}_h = \frac{1}{h}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix},$$

which is the famous $\begin{pmatrix} -1 & 2 & -1 \end{pmatrix}$ matrix associated with the Laplacian operator.

1.7 An error estimate: optimality

We now assess how accurately our finite element solution $u_h \in \mathcal{V}_h$ to (1.12) (or equivalently (1.13)) approximates the solution $u \in \mathcal{V}$ to (1.7) (or equivalently (1.9)). To this end, we need to first define a norm with which we can measure the "closeness" of the approximation. In particular, we introduce the *energy norm* associated with the model problem,

$$|||v||| \equiv \sqrt{a(v,v)} = \left(\int_\Omega \left(\frac{dv}{dx}\right)^2 dx\right)^{1/2} \qquad \forall v \in \mathcal{V}.$$

We here omit the proof, but we can readily show that $a(\cdot,\cdot)$ is an inner product in $\mathcal{V}$, and $|||\cdot|||$ is the associated induced norm. As a consequence, the energy norm of the error $|||u-u_h|||$ is 0 if and only if $u - u_h = 0$.

We next state a key ingredient of the FE error estimate: *Galerkin orthogonality*: since $\ell(v) = a(u, v)$, $\forall v \in \mathcal{V}$, the FE variational statement (1.12) implies

$$0 = \ell(v) - a(u_h, v) = a(u, v) - a(u_h, v) = a(u - u_h, v) = 0 \quad \forall v \in \mathcal{V}_h;$$

the relationship is called Galerkin orthogonality because it states that the error $u - u_h$ is orthogonal to the space $\mathcal{V}_h$ in the inner product $a(\cdot, \cdot)$. We now observe that, for any $w_h \in \mathcal{V}_h$,

$$\begin{aligned}
|||u - u_h|||^2 &= a(u - u_h, u - u_h) && \text{(definition of } |||\cdot|||) \\
&= a(u - u_h, u - w_h) + \cancel{a(u - u_h, w_h - u_h)} && \text{(Galerkin orthogonality)} \\
&= a(u - u_h, u - w_h) \\
&\leq |||u - u_h|||\,|||u - w_h|||. && \text{(Cauchy-Schwarz inequality).}
\end{aligned}$$

We divide both sides by $|||u - u_h|||$ to obtain

$$|||u - u_h||| \leq |||u - w_h||| \quad \forall w_h \in \mathcal{V}_h, \tag{1.17}$$

or, equivalently,

$$|||u - u_h||| = \inf_{w_h \in \mathcal{V}_h} |||u - w_h|||.$$

We observe that the FE approximation $u_h \in \mathcal{V}_h$ is *optimal* in the energy norm in the sense that it is the closest approximation to the solution $u \in \mathcal{V}$ out of all elements in $\mathcal{V}_h$. In other words, even if we knew the solution u, we could not have found a better solution in $\mathcal{V}_h$ than u_h.

As the optimality statement (1.17) holds for any $w_h \in \mathcal{V}_h$, we can set $w_h = \mathcal{I}_h u = \sum_{i=1}^N u(x_i)\phi_i$, the polynomial interpolant associated with our Lagrange basis functions. It can be shown that for any $v \in \mathcal{V}$ the following interpolation error bounds hold:

$$|||v - \mathcal{I}_h v||| \leq h \max_{x \in \Omega} |v''(x)|.$$

(We will later derive the bounds in a more formal setting.) We hence arrive at the following FE error bound in terms of the discretization parameter h:

$$|||u - u_h||| \leq |||u - \mathcal{I}_h u||| \leq h \max_{x \in \Omega} |u''(x)|.$$

In words, the energy-norm error of our finite element solution $u_h \in \mathcal{V}_h$ depends on (i)the maximum second derivative over the domain and (ii) the triangulation parameter h.

To demonstrate the convergence of the finite element approximation, we consider Poisson's problem (1.1) for $f = 1$. The exact solution is $u(x) = x(1 - x)/2$. Figure (1.4) shows that the energy norm of the error $|||u - u_h|||$ converges at the rate of h^1 as predicted by theory.

1.8 Summary

In this lecture, we considered the variational formulation and the associated finite element approximation of one-dimensional Poisson's equation to introduce the main ideas without the complexities associated with higher dimensions and more general equations. We summarize key points of the lecture:

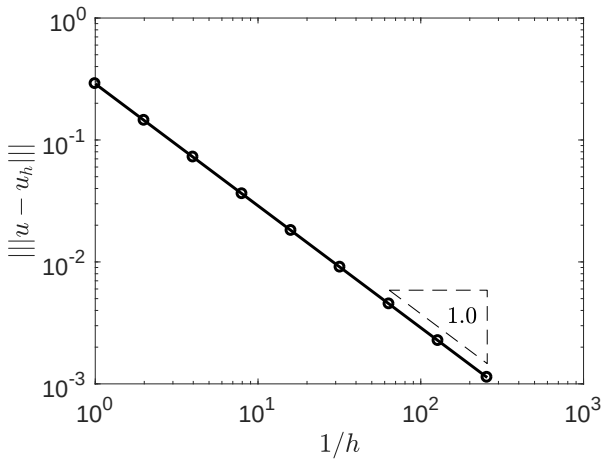


Figure 1.4: Convergence of the finite element approximation for $-\Delta u = 1$.

1. A one-dimensional Poisson's problem can be written in the strong form, minimization form, or variational (or weak) form.

2. The solution to the strong form is also the solution to the minimization and variational form, but the converse is not true in general. The minimization and variational forms admit more general loads f and the associated solutions.

3. A finite element approximation space $\mathcal{V}_h$ is a subspace of $\mathcal{V}$. In one dimension, we may choose $\mathcal{V}_h$ as the space of piecewise linear polynomials associated with a triangulation of $\Omega \subset \mathbb{R}^1$ into segments.

4. The finite element solution is the solution of the minimization or variational problem in the finite element subspace $\mathcal{V}_h \subset \mathcal{V}$.

5. Given a basis for $\mathcal{V}_h$, the finite element solution can be computed in a systematic manner by assembling the associated stiffness matrix and load vector and then solving the resulting linear system.

6. For Poisson's equation, the finite element approximation u_h is optimal in energy norm; even *if* we knew the exact solution $u \in \mathcal{V}$, we could not have found a better solution in $\mathcal{V}_h$ than u_h.

7. The error in the linear finite element approximation of Poisson's equation converges as h^1 in energy norm, where h is the mesh spacing.

Lecture 2

Variational formulation

2.1 Introduction

In the previous lecture, we developed a variational formulation and the associated finite element approximation for one-dimensional Poisson's equation with homogeneous Dirichlet boundary conditions. In this lecture, we focus on the derivation of the variational formulation for problems (i) in higher spatial dimensions, (ii) with more general boundary conditions, and (iii) governed by more general equations. In addition, we discuss the well-posedness of the variational formulation.

2.2 Hilbert and Banach spaces

We start the section with an apology: it may be difficult to appreciate the formalism provided in this section and the next two at this point. But we introduce these spaces upfront such that we can state our weak formulations in a proper functional setting. We will later see that this formalism allows us to provide various theoretical results about the weak formulation and the associated finite element approximation; this theoretical foundation is a strength of the finite element method.

The solutions to the PDEs are most naturally sought in Hilbert spaces. By way of preliminaries, we recall the definition of a *linear space*, *norm*, and *inner product*. We limit ourselves to spaces of real-valued functions; however the following statements readily extend to spaces of complex-valued functions.

Definition 2.1 (linear space)**.** $\mathcal{V}$ is a linear space if the following conditions hold.

Closure axioms

1. Closure under addition: if $w, v \in \mathcal{V}$, then $w + v \in \mathcal{V}$.
2. Closure under scalar multiplication: if $w \in \mathcal{V}$ and $\alpha \in \mathbb{R}$, then $\alpha w \in \mathcal{V}$.

Axioms of addition

3. Commutativity: for all $w, v \in \mathcal{V}$, $w + v = v + w$.

4. Associativity: for all $w, v, z \in \mathcal{V}$, $(w + v) + z = w + (v + z)$.

5. Existence of zero element: there exists $0 \in \mathcal{V}$ such that $v + 0 = v \ \forall v \in \mathcal{V}$.

6. Existence of negatives: for all $v \in \mathcal{V}$, there exists $(-1)v \in \mathcal{V}$ such that $v + (-1)v = 0$.

Axioms of scalar multiplication

7. Associativity: for all $v \in \mathcal{V}$ and $\alpha, \beta \in \mathbb{R}$, $\alpha(\beta v) = (\alpha\beta)v$.

8. Distributivity w.r.t. vector addition: for all $v, w \in \mathcal{V}$ and $\alpha \in \mathbb{R}$, $\alpha(v + w) = \alpha v + \alpha w$.

9. Distributivity w.r.t. scalar addition: for all $v \in \mathcal{V}$ and $\alpha, \beta \in \mathbb{R}$, $(\alpha + \beta)v = \alpha v + \beta v$.

10. Existence of identity: for all $v \in \mathcal{V}$, $1v = v$.

Remark 2.2. If $\mathcal{V}$ is a linear space and $v_1, \ldots, v_n \in \mathcal{V}$, then $\sum_{i=1}^{n} \alpha_i v_i \in \mathcal{V}$ for any $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$.

Definition 2.3 (norm)**.** Given a linear space $\mathcal{V}$, a norm is a function $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ that satisfies the following three conditions: $\forall w, v \in \mathcal{V}$ and $\forall \alpha \in \mathbb{R}$,

1. Absolute scalability: $\|\alpha v\| = |\alpha|\|v\|$;

2. Positive definiteness: $\|v\| \geq 0$, and $\|v\| = 0 \Leftrightarrow v = 0$;

3. Triangle inequality: $\|w + v\| \leq \|w\| + \|v\|$.

Definition 2.4 (inner product)**.** Given a linear space $\mathcal{V}$, an inner product is a function $(\cdot, \cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ that satisfies the following three conditions: $\forall w, v, z \in \mathcal{V}$ and $\forall \alpha, \beta \in \mathbb{R}$,

1. Symmetry: $(w, v) = (v, w)$;

2. Linearity in first argument: $(\alpha w + \beta v, z) = \alpha(w, z) + \beta(v, z)$;

3. Positive definiteness: $(v, v) \geq 0$, and $(v, v) = 0 \Leftrightarrow v = 0$.

Note: the combination of the first and second conditions implies that the inner product is also linear in the second argument.

Definition 2.5 (induced norm)**.** Given a linear space $\mathcal{V}$ and an inner product $(\cdot, \cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, the induced norm $\|\cdot\|$ is given by

$$\|v\| \equiv \sqrt{(v, v)} \quad \forall v \in \mathcal{V}.$$

Remark 2.6 (induced norm)**.** The induced norm is a norm; i.e., it satisfies all three properties of a norm. The absolute scalability follows from linearity:

$$\|\alpha v\|^2 = (\alpha v, \alpha v) = \alpha^2 (v, v) = |\alpha|^2 \|v\|^2.$$

The positive definiteness of the induced norm is a direct consequence of the positive definiteness of the associated inner product. The triangle inequality is proved using the Cauchy-Schwarz inequality in Proposition 2.7: $\forall w, v \in \mathcal{V}$,

$$\|w + v\|^2 = (w + v, w + v) = \|w\|^2 + 2(w, v) + \|v\|^2 \leq \|w\|^2 + 2\|w\|\|v\| + \|v\|^2 = (\|w\| + \|v\|)^2$$

and hence $\|w + v\| \leq \|w\| + \|v\|$.

Proposition 2.7 (Cauchy-Schwarz inequality)**.** Given a linear space $\mathcal{V}$ and an inner product $(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, the associated induced norm $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ satisfies

$$(w, v) \leq \|w\|\|v\| \quad \forall w, v \in \mathcal{V}.$$

Proof. For $\|v\| = 0$, the proof is trivial. For $\|v\| \neq 0$, we observe

$$0 \leq \left\| w - \frac{(w,v)}{\|v\|^2} v \right\|^2 = \|w\|^2 - 2\frac{(w,v)^2}{\|v\|^2} + \frac{(w,v)^2}{\|v\|^2} = \|w\|^2 - \frac{(w,v)^2}{\|v\|^2};$$

the multiplication by $\|v\|^2$ yields $(w, v)^2 \leq \|w\|^2\|v\|^2$ or, equivalently, $(w, v) \leq \|w\|\|v\|$. □

We now define a *Hilbert space* and a *Banach space*.

Definition 2.8 (Hilbert space)**.** A Hilbert space $\mathcal{V}$ is a complete linear space endowed with an inner product $(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and the associated induced norm $\|\cdot\| : \mathcal{V} \to \mathbb{R}$

Definition 2.9 (Banach space)**.** A Banach space $\mathcal{V}$ is a complete linear space endowed with a norm $\|\cdot\| : \mathcal{V} \to \mathbb{R}$.

A space $\mathcal{V}$ is said to be *complete* if any Cauchy sequence with respect to the norm $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ converges to an element of $\mathcal{V}$. A sequence v_1, v_2, v_3, ... is said to be a Cauchy sequence if for any $\delta > 0$ there exists a number N such that $\|v_i - v_j\| \leq \delta$, $\forall i, j > N$. Moreover, the sequence v_i is said to converge to v if $\|v - v_i\| \to 0$ as $i \to \infty$. The readers unfamiliar with the concept of completeness may think of a Hilbert space and a Banach space simply as an inner product space and a normed space. However, completeness is an important property of the Hilbert and Banach spaces, which makes the spaces suitable for the weak formulation of PDEs.

2.3 Sobolev spaces: $L^2(\Omega)$, $H^1(\Omega)$, and $H^1_0(\Omega)$

We now introduce some Hilbert spaces that are most commonly used in the weak formulation of PDEs. By way of preliminaries, we recall the definition of Lipschitz continuous functions.

Definition 2.10 (Lipschitz continuous functions)**.** A function $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous if there exists a Lipschitz constant $L < \infty$ such that

$$|f(x) - f(y)| \leq L|x - y| \quad \forall x, y \in \mathbb{R}.$$

We now characterize the domain Ω with which the function spaces are associated.

Definition 2.11 (Lipschitz domain)**.** A domain $\Omega \subset \mathbb{R}^d$ is called a Lipschitz domain if its boundary $\partial\Omega$ is a graph of a Lipschitz continuous function: corners are permitted, but cusps are not.

In words, Lipschitz domains are domains with a sufficient regular boundary. We will work exclusively with Lipschitz domains in this lecture.

We now introduce a space of square integrable functions on $\Omega \subset \mathbb{R}^d$ (in the Lebesgue sense).

Definition 2.12 ($L^2(\Omega)$ space)**.** The Lebesgue space $L^2(\Omega)$ is endowed with an inner product

$$(w,v)_{L^2(\Omega)} \equiv \int_\Omega wv dx$$

and the associated induced norm $\|w\|_{L^2(\Omega)} \equiv \sqrt{(w,w)_{L^2(\Omega)}}$; the space consists of functions

$$L^2(\Omega) \equiv \{w \mid \|w\|_{L^2(\Omega)} < \infty\}.$$

The $L^2(\Omega)$ space contains functions that are square integrable over Ω, including functions that are discontinuous and also unbounded. For example, consider $x^{-1/4}$ over $\Omega \equiv (0,1)$; the function is unbounded at $x = 0$ but is square integrable and hence is in $L^2(\Omega)$. Two functions in $L^2(\Omega)$ which differ over a set of measure zero — any points in $\mathbb{R}^1$, curves in $\mathbb{R}^2$, and surfaces in $\mathbb{R}^3$ — are deemed equivalent. For instance, consider two functions on $\Omega \equiv (-1,1)$,

$$f(x) \equiv \begin{cases} -1, & x \le 0 \\ 1, & x > 0 \end{cases} \qquad \text{and} \qquad g(x) \equiv \begin{cases} -1, & x < 0 \\ 1, & x \ge 0 \end{cases}.$$

These two functions are equivalent in $L^2(\Omega)$. (We also readily confirm that each function is square integrable.) More formally, the $L^2(\Omega)$ norm of the difference in the two functions is zero; we appeal to the properties of the Lebesgue integration — we can omit any point (or more generally a set of measure zero) — to obtain

$$\|f-g\|^2_{L^2(\Omega)} \equiv \int_{-1}^{1} (f-g)^2 dx = \lim_{\epsilon \to 0} \left(\int_{-1}^{-\epsilon} (\underbrace{f-g}_{=0})^2 dx + \int_{\epsilon}^{1} (\underbrace{f-g}_{=0})^2 dx \right) = 0.$$

Since $\|f-g\|_{L^2(\Omega)} = 0$, f and g are equivalent in $L^2(\Omega)$.

Definition 2.13 (weak derivative in $\mathbb{R}^1$)**.** Let $\Omega \subset \mathbb{R}^1$ be a bounded domain and $C_0^\infty(\Omega)$ be the space of infinitely differentiable functions over Ω whose value and all derivatives are zero at the endpoints. A function $D^1 g : \Omega \to \mathbb{R}$ is a weak first derivative of $g : \Omega \to \mathbb{R}$ if

$$\int_\Omega v D^1 g dx = -\int_\Omega \frac{dv}{dx} g dx \quad \forall v \in C_0^\infty(\Omega).$$

More generally, $D^k g : \Omega \to \mathbb{R}$ is a weak k-th derivative of g if

$$\int_\Omega v D^k g dx = (-1)^k \int_\Omega \frac{d^k v}{dx^k} g dx \quad \forall v \in C_0^\infty(\Omega).$$

To make the idea of weak derivative concrete, consider the absolute-value function $g(x) = |x|$ over $\Omega \equiv (-1,1)$. The function is not differentiable in the classical sense due to the presence of the "kink" at $x = 0$. However, we can readily obtain a weak first derivative $D^1 g$. We wish to find $D^1 g$ such that, $\forall v \in C_0^\infty(\Omega)$,

$$\begin{aligned}
\int_{-1}^{1} v D^1 g dx &= -\int_{-1}^{1} \frac{dv}{dx} g dx = -\lim_{\epsilon \to 0} \left(\int_{-1}^{-\epsilon} \frac{dv}{dx} g dx + \int_{\epsilon}^{1} \frac{dv}{dx} g dx \right) \\
&= -\lim_{\epsilon \to 0} \left(-\int_{-1}^{-\epsilon} v \underbrace{\frac{dg}{dx}}_{-1} dx + \underbrace{[vg]_{x=-1}^{-\epsilon}}_{v(-\epsilon)g(-\epsilon)} - \int_{\epsilon}^{1} v \underbrace{\frac{dg}{dx}}_{1} dx + \underbrace{[vg]_{x=\epsilon}^{1}}_{-v(\epsilon)g(\epsilon)} \right) \\
&= \int_{-1}^{0} -1 v dx + \int_{0}^{1} 1 v dx.
\end{aligned}$$

We observe that a Heaviside-like function

$$(D^1 g)(x) = \begin{cases} -1, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

satisfies the relationship. (The particular value at $x = 0$ is irrelevant because it is a set of measure zero.) Since $D^1 g$ is square integrable, the weak first derivative $D^1 g$ is in $L^2(\Omega)$.

We can repeat the procedure to find a weak second derivative $D^2 g$. We observe that, $\forall v \in C_0^\infty(\Omega)$,

$$\int_{-1}^{1} v D^2 g dx = \int_{-1}^{1} \frac{d^2 v}{dx^2} g dx = \int_{-1}^{0} 1 \frac{dv}{dx} dx + \int_{0}^{1} -1 \frac{dv}{dx} dx = [v]_{x=-1}^{0} - [v]_{x=0}^{1} = 2v(0).$$

Since $2v(0) = \int_{-1}^{1} 2\delta(x) dx$ where δ is the Dirac delta, we find $D^2 g = 2\delta$. However, since the Dirac delta is not square integrable, $D^2 g$ is not in $L^2(\Omega)$.

We now generalize the weak derivative to functions in $\mathbb{R}^d$, $d > 1$.

Definition 2.14 (weak derivative in $\mathbb{R}^d$)**.** Let $\Omega \subset \mathbb{R}^d$ and $C_0^\infty(\Omega)$ be the space of infinitely differentiable functions over Ω whose value and all derivatives are zero on the boundary $\partial\Omega$. A function $\frac{\partial g}{\partial x_i} : \Omega \to \mathbb{R}$, $i = 1, \ldots, d$, is a weak first partial derivative of $g : \Omega \to \mathbb{R}$ if

$$\int_\Omega v \frac{\partial g}{\partial x_i} dx = -\int_\Omega \frac{\partial v}{\partial x_i} g dx \quad \forall v \in C_0^\infty(\Omega).$$

The associated gradient is the vector-valued function $\nabla v \equiv (\frac{\partial v}{\partial x_1}, \ldots, \frac{\partial v}{\partial x_d})$.

Having defined the weak derivative, we now define the $H^1(\Omega)$ space:

Definition 2.15 ($H^1(\Omega)$ space)**.** The Sobolev space $H^1(\Omega)$ is endowed with an inner product

$$(w, v)_{H^1(\Omega)} \equiv \int_\Omega (\nabla v \cdot \nabla w + vw) dx = (\nabla v, \nabla w)_{L^2(\Omega)} + (v, w)_{L^2(\Omega)},$$

and the associated induced norm $\|w\|_{H^1(\Omega)} \equiv \sqrt{(w, w)_{H^1(\Omega)}}$; the space consists of functions

$$H^1(\Omega) \equiv \{w \mid \|w\|_{H^1(\Omega)} < \infty\}.$$

In words, the $H^1(\Omega)$ space consists of functions that are square integrable and whose weak first derivatives are square integrable. For instance, the absolute-value function $g(x) = |x|$ on $\Omega \equiv (-1, 1)$ is in $H^1(\Omega)$ because the function is square integrable and its weak derivative — which is a Heaviside-like function as shown earlier — is square integrable. On the other hand, the Heaviside-like function is not in $H^1(\Omega)$ because its weak first derivative — which is a Dirac delta — is not square integrable. In general, $H^1(\Omega) \subset L^2(\Omega)$ because $H^1(\Omega)$ functions must have a square-integrable weak first derivative whereas $L^2(\Omega)$ functions do not.

Another related space that is frequently encountered in the weak formulation of PDEs is the $H_0^1(\Omega)$ space.

Definition 2.16 ($H_0^1(\Omega)$ space)**.** The $H_0^1(\Omega)$ is endowed with the $H^1(\Omega)$ inner product $(w,v)_{H^1(\Omega)} \equiv \int_\Omega (\nabla v \cdot \nabla w + vw)dx$ and consists of functions

$$H_0^1(\Omega) \equiv \{w \in H^1(\Omega) \mid w|_{\partial\Omega} = 0\},$$

where $\partial\Omega$ denotes the boundary of Ω.

The $H_0^1(\Omega)$ space consists of a subset of $H^1(\Omega)$ functions that vanish on the boundary. Note that $H_0^1(\Omega)$ for $\Omega \equiv (0,1) \subset \mathbb{R}^1$ is precisely the space $\mathcal{V}$ we used in Sections 1.3 and 1.4 for the variational and minimization formulations of one-dimensional Poisson's equation with the homogeneous Dirichlet boundary conditions. By construction $H_0^1(\Omega) \subset H^1(\Omega)$ since the $H^1(\Omega)$ space contains functions that do not vanish on the boundary.

We also introduce the $H^1(\Omega)$ *semi-norm*:

Definition 2.17 ($H^1(\Omega)$ semi-norm)**.** The $H^1(\Omega)$ semi-norm is denoted by $|\cdot|_{H^1(\Omega)}$ and is given by

$$|v|_{H^1(\Omega)} \equiv \left(\int_\Omega \nabla v \cdot \nabla v dx\right)^{1/2} = \|\nabla v\|_{L^2(\Omega)} \quad \forall v \in H^1(\Omega).$$

The $H^1(\Omega)$ semi-norm is not a norm on $H^1(\Omega)$. Specifically, a semi-norm in general does not satisfy the positive definiteness condition. For example, consider a function $v = 1$ on $\Omega \equiv (-1,1)$; the function is clearly not zero, but $|v|_{H^1(\Omega)} = \int_\Omega (\frac{dv}{dx})^2 dx = \int_\Omega 0 dx = 0$.

2.4 Sobolev spaces: more general spaces

While we most frequently use Sobolev spaces $L^2(\Omega)$, $H^1(\Omega)$, and $H_0^1(\Omega)$ in weak formulations of second-order PDEs, more general Sobolev spaces are required for higher-order PDEs. We here introduce these more general spaces for completeness. As the results below can be considered a generalization of the particular results in Section 2.3, we will simply state them.

Definition 2.18 (multi-dimensional derivative)**.** Let $\alpha \equiv (\alpha_1, \ldots, \alpha_d)$ be a d-dimensional multi-index of non-negative integers, and define its absolute value by $|\alpha| \equiv \alpha_1 + \cdots + \alpha_d$. The partial derivative operator D^α is given by

$$D^\alpha(\cdot) \equiv \frac{\partial^{|\alpha|}(\cdot)}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

Definition 2.19 ($H^k(\Omega)$ space)**.** For a non-negative integer k, the Sobolev space $H^k(\Omega)$ is endowed with an inner product

$$(w,v)_{H^k(\Omega)} \equiv \sum_{|\alpha| \le k} (D^\alpha w, D^\alpha v)_{L^2(\Omega)}$$

and the associated induced norm $\|w\|_{H^k(\Omega)} \equiv \sqrt{(w,w)_{H^k(\Omega)}}$; the space consists of functions

$$H^k(\Omega) \equiv \{w \mid \|w\|_{H^k(\Omega)} < \infty\}.$$

Definition 2.20 ($H^k(\Omega)$ semi-norm)**.** The $H^k(\Omega)$ semi-norm is denoted by $|\cdot|_{H^k(\Omega)}$ and is given by

$$|v|_{H^k(\Omega)} \equiv \|D^\alpha v\|_{L^2(\Omega)} \quad \forall v \in H^k(\Omega).$$

Definition 2.21 ($L^p(\Omega)$ space)**.** The Banach space $L^p(\Omega)$ is endowed with a norm

$$\|w\|_{L^p(\Omega)} \equiv \left(\int_\Omega |w|^p dx \right)^{1/p}$$

in the case $1 \leq p < \infty$ and

$$\|w\|_{L^\infty(\Omega)} \equiv \operatorname*{ess\,sup}_{x \in \Omega} |w(x)|$$

in the case $p = \infty$. In either case, the $L^p(\Omega)$ space consists of functions

$$L^p(\Omega) \equiv \{w \mid \|w\|_{L^p(\Omega)} < \infty\}.$$

Definition 2.22 (W_p^k space)**.** The Sobolev space W_p^k is endowed with a norm

$$\|w\|_{W_p^k(\Omega)} \equiv \left(\sum_{|\alpha| \leq k} \|D^\alpha w\|_{L^p(\Omega)}^p \right)^{1/p}$$

in the case $1 \leq p < \infty$ and

$$\|w\|_{W_\infty^k(\Omega)} \equiv \max_{|\alpha| \leq k} \|D^\alpha w\|_{L^\infty(\Omega)}$$

in the case $p = \infty$. In either case, the $W_p^k(\Omega)$ space consists of functions

$$W_p^k(\Omega) = \{w \mid \|w\|_{W_p^k(\Omega)} < \infty\}.$$

Remark 2.23. The $H^k(\Omega)$ space is a special case of $W_p^k(\Omega)$ space for $p = 2$.

2.5 d-dimensional Poisson's problem: homogeneous Dirichlet BC

We consider Poisson's equation in $\mathbb{R}^d$ for $d \geq 1$. To this end, we first introduce a d-dimensional Lipschitz domain $\Omega \subset \mathbb{R}^d$. The strong form of Poisson's equation with homogeneous Dirichlet boundary conditions is as follows: find u such that

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2.1}$$

Here, the Laplacian of u is given by $\Delta u \equiv \frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_d^2}$. Poisson's equation models, for instance, steady heat transfer, where u is the temperature field (relative to the ambient temperature), f is the volume heat source, and the homogeneous Dirichlet boundary condition corresponds to the fixed-temperature condition.

The variational formulation of (2.1) requires an appropriate choice of a function space. For homogeneous Dirichlet boundary condition, the appropriate Sobolev space is

$$\mathcal{V} \equiv H_0^1(\Omega).$$

We recall that $H_0^1(\Omega) \equiv \{w \in H^1(\Omega) \mid w|_{\partial\Omega} = 0\}$; i.e., the space consists of functions (i) whose value and first derivative are square integrable and (ii) that vanish on the boundary. Note that any function $w \in H_0^1(\Omega)$ satisfies the boundary condition $u|_{\partial\Omega} = 0$ by construction.

To obtain a variational (or weak) form, we employ the weighted residual method: we multiply (2.1) by a test function $v \in \mathcal{V}$, integrate the expression, and then integrate by parts the left hand side:

$$\int_\Omega v(-\Delta u)dx = \int_\Omega vf dx \quad \Rightarrow \quad \int_\Omega \nabla v \cdot \nabla u dx - \underbrace{\int_{\partial\Omega} v\frac{\partial u}{\partial n}ds}_{=0} = \int_\Omega vf dx;$$

the boundary term vanishes because $v|_{\partial\Omega} = 0$ for $v \in \mathcal{V} \equiv H_0^1(\Omega)$. We now recognize the bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$,

$$a(w, v) \equiv \int_\Omega \nabla v \cdot \nabla w dx \quad \forall w, v \in \mathcal{V},$$

and the linear form $\ell : \mathcal{V} \to \mathbb{R}$,

$$\ell(v) \equiv \int_\Omega vf dx \quad \forall v \in \mathcal{V}.$$

Our variational problem is as follows: find $u \in \mathcal{V}$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathcal{V}. \tag{2.2}$$

Using exactly the same procedure as the one-dimensional case shown in Section 1.3, we can show that the solution to the strong form (2.1) satisfies the variational form (2.2). However the converse is not true in general; the variational form admits more general loads f and hence solutions than the strong form.

2.6 Mixed problems: essential and natural boundary conditions

We have so far considered Poisson's equation with a homogeneous Dirichlet boundary condition. We now consider a problem with a mixed boundary condition. To this end, given $\Omega \subset \mathbb{R}^d$, we first partition the domain boundary $\partial\Omega$ into a Dirichlet part Γ_D and a Neumann part Γ_N such that $\Gamma_D \cap \Gamma_N = \emptyset$, $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, and $\Gamma_D \neq \emptyset$. The first two conditions ensure that there is one and only one boundary condition (Dirichlet or Neumann) imposed at any part of the boundary $\partial\Omega$; the last condition is required to ensure the problem is well-posed, as we will see later. We then consider the following boundary value problem: find u such that

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_D, \\ \frac{\partial u}{\partial n} &= g \quad \text{on } \Gamma_N, \end{aligned} \tag{2.3}$$

where f is the volume source term and g is the boundary source term. In the case of a steady heat transfer, f and g represent volume and boundary heat sources, respectively.

To obtain a variational form of (2.3), we modify the function space from the homogeneous Dirichlet boundary condition case. The function space suitable for the mixed boundary condition case is

$$\mathcal{V} \equiv \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}. \tag{2.4}$$

Note that $H_0^1(\Omega) \subset \mathcal{V} \subset H^1(\Omega)$; functions in $H_0^1(\Omega)$ must vanish everywhere on $\partial\Omega$, functions in $\mathcal{V}$ must vanish only on $\Gamma_D \subset \partial\Omega$, and functions in $H^1(\Omega)$ have no conditions on their boundary

values. We now apply the weighted residual method to obtain the variational form: we multiply (2.3) by a test function $v \in \mathcal{V}$, integrate the expression, and then integrate by parts the left hand side:

$$\int_\Omega v(-\Delta u)dx = \int_\Omega vf dx \quad \Rightarrow \quad \int_\Omega \nabla v \cdot \nabla u dx - \underbrace{\int_{\Gamma_D} v\frac{\partial u}{\partial n}ds}_{=0} - \int_{\Gamma_N} v\frac{\partial u}{\partial n}ds = \int_\Omega vf dx;$$

the boundary term on Γ_D vanishes because $v|_{\Gamma_D} = 0$ for $v \in \mathcal{V}$. On the other hand, the boundary term on Γ_N remains. We now replace $\frac{\partial u}{\partial n}$ with g to incorporate the boundary condition we wish to impose: $\frac{\partial u}{\partial n} = g$ on Γ_N. The resulting weighted residual form is

$$\int_\Omega \nabla v \cdot \nabla u dx = \int_\Omega vf dx + \int_{\Gamma_N} vg ds.$$

We now recognize the bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ given by

$$a(w, v) \equiv \int_\Omega \nabla v \cdot \nabla w dx \quad \forall w, v \in \mathcal{V},$$

and the linear form $\ell : \mathcal{V} \to \mathbb{R}$ given by

$$\ell(v) \equiv \int_\Omega vf dx + \int_{\Gamma_N} vg ds \quad \forall v \in \mathcal{V}.$$

Our variational problem is as follows: find $u \in \mathcal{V}$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathcal{V}. \tag{2.5}$$

We readily observe that the solution to the strong form (2.3) satisfies the variational form (2.5); for all $v \in \mathcal{V}$,

$$\begin{aligned} a(u, v) - \ell(v) &\equiv \int_\Omega \nabla v \cdot \nabla u dx - \int_\Omega vf dx - \int_{\Gamma_N} vg ds \\ &= \int_\Omega v(-\Delta u)dx + \int_{\Gamma_D} v\frac{\partial u}{\partial n}ds + \int_{\Gamma_N} v\frac{\partial u}{\partial n}ds - \int_\Omega vf dx - \int_{\Gamma_N} vg ds \\ &= \int_\Omega v \underbrace{(-\Delta u - f)}_{=0 \text{ as } -\Delta u = f \text{ in } \Omega} dx + \underbrace{\int_{\Gamma_D} v\frac{\partial u}{\partial n}ds}_{=0 \text{ as } v|_{\Gamma_D} = 0} + \int_{\Gamma_N} v \underbrace{\left(\frac{\partial u}{\partial n} - g\right)}_{=0 \text{ as } \frac{\partial u}{\partial n} = g \text{ on } \Gamma_N} ds = 0. \end{aligned}$$

Hence a solution to the strong form (2.3) is a solution to the variational form (2.5); however, again the converse is not true as the variational form admits more general forms of f and g than the strong form.

In the variational formulation of the mixed boundary condition, the Dirichlet and Neumann conditions are treated differently. On one hand, we explicitly impose the Dirichlet boundary condition $u = 0$ on Γ_D through the choice of the space $\mathcal{V}$ in (2.4). On the other hand, the Neumann boundary condition $\frac{\partial u}{\partial n} = g$ on Γ_N is implicitly contained in the variational statement (2.5). A boundary condition that is explicitly imposed by the choice of the function space is called an *essential boundary condition*; a boundary condition that is implicitly imposed by the variational statement is called a *natural boundary condition*. In the above treatment of Poisson's equation with mixed boundary conditions, the Dirichlet condition is an essential boundary condition, and the Neumann condition is a natural boundary condition.

2.7 Nonhomogeneous Dirichlet boundary condition

We now consider a problem with a *nonhomogeneous* Dirichlet boundary condition. The strong form is as follows: find u such that

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= u^B \quad \text{on } \Gamma_D \equiv \partial\Omega \end{aligned} \tag{2.6}$$

for some boundary function u^B and source term f. While we here focus on the pure Dirichlet problem for simplicity, the approach in this section can be combined with the approach for mixed problems in Section 2.6 to treat mixed problems with nonhomogeneous Dirichlet and Neumann boundary conditions.

To obtain a variational form of (2.6), we introduce spaces

$$\begin{aligned} \mathcal{V}^E &\equiv \{w \in H^1(\Omega) \mid w|_{\Gamma_D} = u^B\}, \\ \mathcal{V} &\equiv H_0^1(\Omega). \end{aligned}$$

The superscript "E" stands for "essential", as the space $\mathcal{V}^E$ satisfies the essential (i.e., Dirichlet) boundary condition. Note that, for $u^B \neq 0$, $\mathcal{V}^E$ is *not* a linear space; for $w, v \in \mathcal{V}^E$, $z = w + v \notin \mathcal{V}^E$ because $z|_{\Gamma_D} = 2u^B \neq u^B$. Rather, $\mathcal{V}^E$ is an *affine space*: given an arbitrary fixed element $u^E \in \mathcal{V}^E$ so that $u^E|_{\Gamma_D} = u^B$, we have $\mathcal{V}^E = u^E + \mathcal{V} = \{u^E + v \mid v \in \mathcal{V}\}$. We now employ the weighted residual method: we multiply (2.6) by a test function v in the *linear space* $\mathcal{V}$ — and not *affine space* $\mathcal{V}^E$ —, integrate the expression, and integrate by parts the right hand side to obtain

$$\int_\Omega v(-\Delta u)dx = \int_\Omega vf dx \quad \Rightarrow \quad \int_\Omega \nabla v \cdot \nabla u dx - \underbrace{\int_{\partial\Omega} v\frac{\partial u}{\partial n}ds}_{=0} = \int_\Omega vf dx;$$

again the boundary term vanishes because v is in $\mathcal{V}$ (and not $\mathcal{V}^E$). We recognize a bilinear form and a linear form

$$\begin{aligned} a(w, v) &\equiv \int_\Omega \nabla v \cdot \nabla w dx \quad \forall w, v \in \mathcal{V}, \\ \ell(v) &\equiv \int_\Omega vf dx \quad \forall v \in \mathcal{V}. \end{aligned}$$

The variational problem is as follows: find $u \in \mathcal{V}^E$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathcal{V}. \tag{2.7}$$

We note that the bilinear form and linear form are identical to the homogeneous Dirichlet boundary condition case considered in Section 2.5. However, our trial space is different; the space $\mathcal{V}^E$ is an affine space of functions that satisfy the nonhomogeneous Dirichlet boundary condition. As discussed in Section 2.6, a Dirichlet boundary condition is an essential boundary condition, which is explicitly imposed through the choice of the space.

We readily observe that the solution to the strong form (2.6) satisfies the variational form (2.7); for all $v \in \mathcal{V}$,

$$\begin{aligned} a(u,v) - \ell(v) &\equiv \int_\Omega \nabla v \cdot \nabla u dx - \int_\Omega v f dx - \int_{\Gamma_N} v g ds \\ &= \int_\Omega v(-\Delta u) dx + \int_{\partial\Omega} v \frac{\partial u}{\partial n} ds - \int_\Omega v f dx \\ &= \int_\Omega v \underbrace{(-\Delta u - f)}_{=0 \text{ as } -\Delta u = 0 \text{ in } \Omega} dx + \underbrace{\int_{\partial\Omega} v \frac{\partial u}{\partial n} ds}_{=0 \text{ as } v|_{\Gamma_D \equiv \partial\Omega} = 0}. \end{aligned}$$

Moreover, the boundary condition $u = u^B$ on $\Gamma_D \equiv \partial\Omega$ is satisfied because $u \in \mathcal{V}^E$. Hence a solution to the strong form (2.6) is a solution to the variational form (2.7); however, again the converse is not true as the variational form admits more general solutions.

In practice, it is more convenient to reformulate the problem such that both the trial and test spaces are linear. We first choose an arbitrary fixed function u^E in $\mathcal{V}^E$ that satisfies the nonhomogeneous Dirichlet boundary condition so that $u^E|_{\Gamma_D} = u^B$. We then express the solution u as $u = u^E + \tilde{u}$ for $\tilde{u}$ in the linear space $\mathcal{V}$, and rearrange the variational form (2.7) as

$$a(u^E + \tilde{u}, v) = \ell(v) \quad \Rightarrow \quad a(\tilde{u}, v) = \ell(v) - a(u^E, v).$$

We now recognize the right hand side $\ell(\cdot) - a(u^E, \cdot)$ as another linear form on $\mathcal{V}$ and formally introduce $\tilde{\ell} : \mathcal{V} \to \mathbb{R}$ such that

$$\tilde{\ell}(v) \equiv \ell(v) - a(u^E, v) \quad \forall v \in \mathcal{V}.$$

We then consider a variational problem for $\tilde{u}$: find $\tilde{u} \in \mathcal{V}$ such that

$$a(\tilde{u}, v) = \tilde{\ell}(v) \quad \forall v \in \mathcal{V}. \tag{2.8}$$

Once we find $\tilde{u}$, we then set $u = u^E + \tilde{u}$, which is in $\mathcal{V}^E$. Note that $\tilde{u} \in \mathcal{V}$ depends on our choice of $u^E \in \mathcal{V}^E$ because $\tilde{\ell}(\cdot)$ depends on u^E; however, the actual solution $u = u^E + \tilde{u}$ is independent of the particular choice of $u^E \in \mathcal{V}^E$.

2.8 General second-order elliptic equation

We have so far considered the variational formulation of Poisson's equation with various boundary conditions. We can readily extend our approach to treat general second-order elliptic equations. To demonstrate the idea, we consider a convection-reaction-diffusion equation with (nonhomogeneous) Dirichlet, Neumann, and Robin boundary conditions. To this end, we partition the Lipschitz domain $\Omega \subset \mathbb{R}^d$ into the Dirichlet boundary Γ_D, the Neuamnn boundary Γ_N, and the Robin boundary Γ_R such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N \cup \overline{\Gamma}_R$; we assume $\Gamma_D \cup \Gamma_R \neq \emptyset$. We then consider a problem of the following form: find u such that

$$\begin{aligned} -\nabla \cdot (\kappa \nabla u) + b \cdot \nabla u + cu &= f \quad \text{in } \Omega \\ u &= u^B \quad \text{on } \Gamma_D \\ n \cdot \kappa \nabla u &= g \quad \text{on } \Gamma_N \\ n \cdot \kappa \nabla u + du &= q \quad \text{on } \Gamma_R, \end{aligned} \tag{2.9}$$

where $\kappa : \Omega \to \mathbb{R}^{d\times d}$ is the diffusivity tensor, $b : \Omega \to \mathbb{R}^d$ is the advection vector, $c : \Omega \to \mathbb{R}$ is the reaction constant, $f : \Omega \to \mathbb{R}$ is the source term, $n : \partial\Omega \to \mathbb{R}^d$ is the outward-point normal on $\partial\Omega$, $u^B : \Gamma_D \to \mathbb{R}$ is the Dirichlet boundary function, $g : \Gamma_N \to \mathbb{R}$ is the Neumann source term, $d : \Gamma_R \to \mathbb{R}$ is the Robin coefficient, and $q : \Gamma_R \to \mathbb{R}$ is the Robin source term. In general each coefficient is spatially varying and hence is a function of space; e.g., the diffusivity tensor evaluated at $x \in \Omega \subset \mathbb{R}^d$ is in $\mathbb{R}^{d\times d}$ and hence is denoted $\kappa : \Omega \to \mathbb{R}^{d\times d}$. For the second-order PDE to be *elliptic*, we require that the diffusitivity tensor is symmetric positive definite almost everywhere: $\kappa(x) \in \mathbb{R}^{d\times d}$ satisfies

$$\xi^T \kappa(x)\xi > 0 \quad \forall \xi \neq 0 \quad \text{a.e. in } \Omega.$$

Also note that in order for the differentiation $\nabla \cdot (\kappa \nabla u)$ for the *strong formulation* (2.9) to be well defined, the diffusivity tensor field must satisfy certain smoothness conditions; we will soon see that this is not a requirement for the weak formulation.

To obtain a variational form of (2.9), we introduce spaces

$$\mathcal{V}^E \equiv \{w \in H^1(\Omega) \mid w|_{\Gamma_D} = u^B\},$$
$$\mathcal{V} \equiv \{w \in H^1(\Omega) \mid w_{\Gamma_D} = 0\}.$$

As discussed in Section 2.7, the Dirichlet boundary condition is imposed strongly; the Neumann and Robin boundary conditions are imposed weakly. We now multiply (2.9) by a test function v in the linear space $\mathcal{V}$, integrate the expression, and integrate by parts the diffusion term to obtain

$$\int_\Omega v(-\nabla \cdot \kappa \nabla u + b \cdot \nabla u + cu - f)dx = 0$$
$$\Rightarrow \int_\Omega (\nabla v \cdot \kappa \nabla u + vb \cdot \nabla u + cvu - vf)dx - \underbrace{\int_{\Gamma_D} vn \cdot \kappa \nabla u ds}_{\text{(D)}} - \underbrace{\int_{\Gamma_N} vn \cdot \kappa \nabla u ds}_{\text{(N)}} - \underbrace{\int_{\Gamma_R} vn \cdot \kappa \nabla u ds}_{\text{(R)}}$$

We now impose the boundary conditions. The Dirichlet boundary condition is imposed strongly by the choice of the trial space $\mathcal{V}^E$ and the test space $\mathcal{V}$; the term (D) vanishes because $v|_{\Gamma_D} = 0$ for all $v \in \mathcal{V}$. The Neumann boundary condition $n \cdot \kappa \nabla u = g$ is weakly imposed; we replace the boundary term (N) by $\int_{\Gamma_N} vg ds$. The Robin boundary condition $n \cdot \kappa \nabla u + du = q$ is also weakly imposed; we replace the boundary term (R) by $\int_{\Gamma_R} v(-du + q)ds$. Upon the substitution of the appropriate boundary conditions, our weighted-residual formulation reads as follows: find $u \in \mathcal{V}^E$ such that

$$\int_\Omega (\nabla v \cdot \kappa \nabla u + vb \cdot \nabla u + cvu - vf)dx - \int_{\Gamma_N} vg ds - \int_{\Gamma_R} v(-du + q)ds = 0 \quad \forall v \in \mathcal{V}.$$

Some reorganization of the terms yield the following variational formulation: find $u \in \mathcal{V}^E$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathcal{V}, \tag{2.10}$$

where

$$a(w, v) \equiv \int_\Omega (\nabla v \cdot \kappa \nabla w + vb \cdot \nabla u + cvu)dx + \int_{\Gamma_R} dvw ds \quad \forall w, v \in \mathcal{V},$$
$$\ell(v) \equiv \int_\Omega fv dx + \int_{\Gamma_N} gv ds + \int_{\Gamma_R} qv ds. \quad \forall v \in \mathcal{V}.$$

With the variational formulation (2.10), unlike with the strong formulation (2.9), we need not assume any smoothness of the coefficients κ, b, or c. We only require that the coefficients are bounded: $\kappa \in (L^\infty(\Omega))^{d\times d}$, $b \in (L^\infty(\Omega))^d$, and $c \in L^\infty(\Omega)$.

As discussed in Section 2.7, in practice, the nonhomogeneous Dirichlet boundary conditions are more conveniently treated through the decomposition of the solution as $u = u^E + \tilde{u}$ for some arbitrary but fixed $u^E \in \mathcal{V}^E$ so that $u^E|_{\Gamma_D} = u^B$ and $\tilde{u} \in \mathcal{V}$. We then consider the following variational problem: find $\tilde{u} \in \mathcal{V}$ such that

$$a(\tilde{u}, v) = \tilde{\ell}(v) \quad \forall v \in \mathcal{V},$$

where $\tilde{\ell} : \mathcal{V} \to \mathbb{R}$ is the modified linear form such that $\tilde{\ell}(v) \equiv \ell(v) - a(u^E, v)$, $\forall v \in \mathcal{V}$; we then set $u = u^E + \tilde{u}$.

2.9 Well-posedness of the weak formulation

We now address a fundamental question: what conditions should a weak formulation satisfy to guarantee the existence *and* uniqueness of the solution? We recall that, in general, a weak formulation is defined by a test space $\mathcal{V}$, trial space $\mathcal{V}^E \equiv u^E + \mathcal{V}$, bilinear form $a(\cdot,\cdot)$, and linear form $\ell(\cdot)$; as such, we wish to identify conditions that these ingredients must satisfy to ensure the existence and uniqueness of the solution.

We first provide a few definitions that characterize a linear form and bilinear form.

Definition 2.24 (dual space)**.** The space of linear functionals $\ell : \mathcal{V} \to \mathbb{R}$ is denoted by $\mathcal{V}'$.

Definition 2.25 (dual norm and continuity)**.** The dual norm of a linear functional $\ell \in \mathcal{V}'$ is given by

$$\|\ell\|_{\mathcal{V}'} \equiv \sup_{v\in\mathcal{V}} \frac{|\ell(v)|}{\|v\|_{\mathcal{V}}}.$$

A linear functional is *continuous* on $\mathcal{V}$ if $\|\ell\|_{\mathcal{V}'} < \infty$.

Corollary 2.26. If a linear form $\ell \in \mathcal{V}'$ is continuous so that $\|\ell\|_{\mathcal{V}'} < \infty$, then

$$|\ell(v)| \le \|\ell\|_{\mathcal{V}'}\|v\|_{\mathcal{V}} \quad \forall v \in \mathcal{V}.$$

In other words, a linear form is continuous if $\exists c < \infty$ such that $|\ell(v)| \le c\|v\|_{\mathcal{V}}$, $\forall v \in \mathcal{V}$.

Example 2.27. Let $f \in L^2(\Omega)$, $\ell(v) \equiv \int_\Omega vf dx$ $\forall v \in \mathcal{V}$, and $\|\cdot\|_{\mathcal{V}} \equiv \|\cdot\|_{H^1(\Omega)}$. Then

$$\|\ell\|_{\mathcal{V}'} \equiv \sup_{v\in\mathcal{V}} \frac{|\ell(v)|}{\|v\|_{\mathcal{V}}} = \sup_{v\in\mathcal{V}} \frac{|\int_\Omega vf dx|}{\|v\|_{\mathcal{V}}} \le \sup_{v\in\mathcal{V}} \frac{\|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)}}{\|v\|_{\mathcal{V}}} \le \sup_{v\in\mathcal{V}} \frac{\|f\|_{L^2(\Omega)}\|v\|_{\mathcal{V}}}{\|v\|_{\mathcal{V}}} = \|f\|_{L^2(\Omega)},$$

where the first inequality follows from Cauchy-Schwarz inequality, and the second inequality follows from $\|v\|_{L^2(\Omega)} \le \|v\|_{H^1(\Omega)} = \|v\|_{\mathcal{V}}$. Hence $\|\ell\|_{\mathcal{V}'} \le \|f\|_{L^2(\Omega)} < \infty$, and ℓ is continuous.

Definition 2.28 (continuity)**.** The continuity constant of a bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is given by

$$\gamma \equiv \sup_{w\in\mathcal{V}} \sup_{v\in\mathcal{V}} \frac{|a(w, v)|}{\|w\|_{\mathcal{V}}\|v\|_{\mathcal{V}}}.$$

A bilinear form is *continuous* on $\mathcal{V}$ if $\gamma < \infty$.

Corollary 2.29. A bilinear form is continuous if there exists $\tilde{\gamma} < \infty$ such that

$$|a(w,v)| \leq \tilde{\gamma}\|w\|_{\mathcal{V}}\|v\|_{\mathcal{V}} \quad \forall w, v \in \mathcal{V},$$

as the condition implies that $\gamma \leq \tilde{\gamma} < \infty$.

Definition 2.30 (coercivity)**.** The coercivity constant of a bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is given by

$$\alpha \equiv \inf_{v \in \mathcal{V}} \frac{a(v,v)}{\|v\|_{\mathcal{V}}^2}.$$

A bilinear form is *coercive* on $\mathcal{V}$ if $\alpha > 0$.

Corollary 2.31. A bilinear form is coercive if there exists $\tilde{\alpha} > 0$ such that

$$a(v,v) \geq \tilde{\alpha}\|v\|_{\mathcal{V}}^2 \quad \forall v \in \mathcal{V},$$

as the condition implies that $\alpha \geq \tilde{\alpha} > 0$.

The following theorem provides an answer to the question regarding the existence and uniqueness of a weak solution.

Theorem 2.32 (Lax-Milgram)**.** Given a Hilbert space $\mathcal{V}$, a continuous, coercive bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, and a continuous linear functional $\ell \in \mathcal{V}'$, there exists a unique $u \in \mathcal{V}$ such that

$$a(u,v) = \ell(v) \quad \forall v \in \mathcal{V}. \tag{2.11}$$

Proof. The proof of existence is beyond the scope of this course. We refer to Brenner and Scott (2008).

The proof of uniqueness is as follows. Suppose we have two solutions $u_1 \in \mathcal{V}$ and $u_2 \in \mathcal{V}$ that are distinct ($u_1 \neq u_2$) and satisfy (2.11): i.e., $a(u_1, v) = \ell(v)$, $\forall v \in \mathcal{V}$, and $a(u_2, v) = \ell(v)$, $\forall v \in \mathcal{V}$. The subtraction of the two equations yields $a(u_1, v) - a(u_2, v) = 0$, $\forall v \in \mathcal{V}$. We then invoke bilinearity to obtain $a(u_1 - u_2, v) = 0$, $\forall v \in \mathcal{V}$. We then choose $v = u_1 - u_2$, which yields $a(u_1 - u_2, u_1 - u_2) = 0$. The coercivity of the bilinear form implies that $a(u_1 - u_2, u_1 - u_2) \geq \alpha\|u_1 - u_2\|_{\mathcal{V}}^2$; since the left hand side is 0, we obtain $\|u_1 - u_2\|_{\mathcal{V}}^2 = 0$. Hence, we arrive at the contradiction: $u_1 = u_2$. If two solutions satisfy (2.11), then they must be the same; the solution to (2.11) is unique. □

The Lax-Milgram theorem provides sufficient conditions under which a weak formulation possess a unique solution. We however note that these are only sufficient, and not necessary, conditions. The above proof for uniqueness also shows how certain properties of the ingredients, such as bilinearity and coercivity, are used to prove the desired result.

The Lax-Milgram theorem concerns with the existence and uniqueness of a weak solution. We now introduce a *stability* or *well-posedness* result which shows that the solution u depends continuously on the data $\ell(\cdot)$.

Proposition 2.33 (stability)**.** Suppose the conditions of the Lax-Milgram theorem, Theorem 2.32, are satisfied. Then, the solution u satisfies

$$\|u\|_{\mathcal{V}} \leq \frac{1}{\alpha}\|\ell\|_{\mathcal{V}'},$$

where α is the coercivity constant.

Proof. The proof is trivial for $\|u\|_{\mathcal{V}} = 0$. For $\|u\|_{\mathcal{V}} \neq 0$, we appeal to the coercivity of the bilinear form and the continuity of the linear form:

$$\alpha \|u\|_{\mathcal{V}}^2 \leq a(u,u) = \ell(u) \leq \|\ell\|_{\mathcal{V}'} \|u\|_{\mathcal{V}}.$$

The division by $\|u\|_{\mathcal{V}} \neq 0$ yields the desired result. □

Corollary 2.34. Suppose the conditions of the Lax-Milgram theorem, Theorem 2.32, are satisfied. In addition, consider two linear forms $\ell_1(\cdot)$ and $\ell_2(\cdot)$, and the associated solutions u_1 and u_2. Then, the difference in the solutions, $u_1 - u_2$, is bounded by the difference in the data, $\ell_1 - \ell_2$:

$$\|u_1 - u_2\|_{\mathcal{V}} \leq \frac{1}{\alpha} \|\ell_1 - \ell_2\|_{\mathcal{V}'}.$$

The stability result shows that the energy norm of the solution u is bounded by the dual norm of the data ℓ. The closely related result in the corollary can be interpreted to mean that a small disturbance in the data ℓ results in a small perturbation in the solution u.

Before we conclude this section, we remark on the well-posedness of problems with nonhomogeneous Dirichlet data. The Lax-Milgram theorem 2.32 requires both the test and trial spaces to be a Hilbert space and in particular linear. We however recall that the trial space for a problem with nonhomogeneous Dirichlet data is an affine space $\mathcal{V}^E \equiv u^E + \mathcal{V}$ for some fixed $u^E \in H^1(\Omega)$ and a linear space $\mathcal{V}$. To prove the existence and uniqueness of the solution, we rely on the reformulated variational formulation (2.8), which decomposes the solution $u \in \mathcal{V}^E$ as $u = u^E + \tilde{u}$ for an arbitrary (but fixed u^E) and $\tilde{u} \in \mathcal{V}$. We specifically apply the Lax-Milgram theorem to the problem for $\tilde{u}$: find $\tilde{u} \in \mathcal{V}$ such that

$$a(\tilde{u}, v) = \tilde{\ell}(v) \quad \forall v \in \mathcal{V},$$

where $\tilde{\ell} : \mathcal{V} \to \mathbb{R}$ is the reformulated linear form $\tilde{\ell}(v) \equiv \ell(v) - a(u^E, v)$, $\forall v \in \mathcal{V}$. By the Lax-Milgram theorem, a unique solution to the problem exists if $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is coercive and continuous, and $\tilde{\ell} \in \mathcal{V}'$ is continuous. The latter requires that $\exists C < \infty$ such that $|\tilde{\ell}(v)| \leq C\|v\|_{\mathcal{V}}$, $\forall v \in \mathcal{V}$. We now observe that, assuming $\ell(\cdot)$ and $a(\cdot,\cdot)$ are continuous,

$$|\tilde{\ell}(v)| \equiv |\ell(v) - a(u^E, v)| \leq |\ell(v)| + |a(u^E, v)| \leq c\|v\|_{\mathcal{V}} + \gamma \|u^E\|_{\mathcal{V}} \|v\|_{\mathcal{V}} = (c + \gamma \|u^E\|_{\mathcal{V}}) \|v\|_{\mathcal{V}},$$

where c and γ are the continuity constant for $\ell(\cdot)$ and $a(\cdot,\cdot)$, respectively. Hence $\tilde{\ell}(\cdot)$ is continuous under these assumptions, and we can apply the Lax-Milgram theorem to show the existence and uniqueness of $\tilde{u} \in \mathcal{V}$ and in turn $u = u^E + \tilde{u} \in \mathcal{V}^E$.

2.10 Poincaré-Friedrichs and trace inequalities

The proof of existence and uniqueness of a weak solution using the Lax-Milgram theorem relies on the continuity and coercivity of the bilinear form and the continuity of the linear form. For many bilinear forms associated with boundary value problems, the verification of coercivity relies on the Poincaré-Friedrichs inequality.

Proposition 2.35 (Poincaré-Friedrichs inequality)**.** Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, and suppose $\Gamma \subset \partial\Omega$ and $\Gamma \neq \emptyset$. Then, there exists a constant $C_{\mathrm{PF}} < \infty$ that only depends on Ω and Γ such that

$$\|v\|_{L^2(\Omega)}^2 \leq C_{\mathrm{PF}} (|v|_{H^1(\Omega)}^2 + \|v\|_{L^2(\Gamma)}^2) \quad \forall v \in H^1(\Omega).$$

Proof. Proof of the proposition is beyond the scope of this course. We refer to Brenner and Scott (2008). □

Corollary 2.36. Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, and suppose $\Gamma_D \subset \partial\Omega$ and $\Gamma_D \neq \emptyset$. Let $\mathcal{V} \equiv \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}$. Then, there exists a constant $C_{\mathrm{PF}} < \infty$ that depends only on Ω and Γ_D such that

$$\|v\|^2_{L^2(\Omega)} \leq C_{\mathrm{PF}} |v|^2_{H^1(\Omega)} \quad \forall v \in \mathcal{V}.$$

Proposition 2.35 allows us to bound the $L^2(\Omega)$ norm of a function by the $H^1(\Omega)$ *semi*-norm of the function and the $L^2(\Gamma)$ norm of the trace of the function on a portion of the boundary, $\Gamma \subset \partial\Omega$. Corollary 2.36 is a specialization of the results for functions that vanish on (at least) a portion of the boundary such that $\|u\|_{L^2(\Gamma_D)} = 0$. Intuitively, the corollary bounds the (integrated) value of the function by the (integrated) gradient of the function, assuming that the function is "pinned" to vanish over a portion of the boundary.

Remark 2.37 (Naming of the Poincaré-Friedrichs inequality)**.** Proposition 2.36 and the related results, such as Corollary 2.36, are sometimes called just Poincaré inequality or Friedrichs inequality. In this note, we will refer to inequalities of these types collectively as Poincaré-Friedrichs inequalities.

Remark 2.38. The smallest C_{PF} that satisfies the Poincaré-Friedrichs inequality is given by

$$C_{\mathrm{PF}} = \sup_{v \in H^1(\Omega)} \frac{\|v\|^2_{L^2(\Omega)}}{|v|^2_{H^1(\Omega)} + \|v\|^2_{L^2(\Gamma)}}.$$

By Rayleigh quotient, the constant is the largest eigenvalue of the following eigenproblem: find $(w, \lambda) \in H^1(\Omega) \times \mathbb{R}$ such that

$$\int_\Omega vw dx = \lambda \left(\int_\Omega \nabla v \cdot \nabla w dx + \int_\Gamma vw dx \right) \quad \forall v \in H^1(\Omega);$$

i.e., $C_{\mathrm{PF}} = \sup\{\lambda\}$.

For many bilinear forms associated with boundary value problems, the verification of continuity relies on the trace inequality.

Proposition 2.39 (trace inequality)**.** Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. Then, there exists a constant $C_{\mathrm{tr}} < \infty$ that depends only on Ω such that

$$\|v\|_{L^2(\partial\Omega)} \leq C_{\mathrm{tr}} \|v\|_{H^1(\Omega)} \quad \forall v \in H^1(\Omega).$$

Proof. Proof of the proposition is beyond the scope of this course. We refer to Brenner and Scott (2008). □

The trace inequality is particularly useful when we wish to show continuity of a linear or bilinear form which involves integration over (a part of) the boundary.

Remark 2.40. The smallest $C_{\rm tr}$ that satisfies the trace inequality is given by

$$C_{\rm tr} = \sup_{v \in H^1(\Omega)} \frac{\|v\|_{L^2(\partial\Omega)}}{\|v\|_{H^1(\Omega)}}.$$

By Rayleigh quotient, the constant is the square root of the largest eigenvalue of the following eigenproblem: find $(w, \lambda) \in H^1(\Omega) \times \mathbb{R}$ such that

$$\int_{\partial\Omega} vw dx = \lambda(\int_\Omega \nabla v \cdot \nabla w dx + \int_\Omega vw dx) \quad \forall v \in H^1(\Omega);$$

i.e., $C_{\rm tr} = \sqrt{\sup\{\lambda\}}$.

2.11 Example: well-posedness of Poisson's problem

We now demonstrate the application of the Lax-Milgram theorem to prove that the mixed Poisson's problem considered in Section 2.6 is well-posed. The problem is reproduced here for convenience. Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, Γ_D and Γ_N be Dirichlet and Neumann boundaries such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$ and $\Gamma_D \neq \emptyset$, and $\mathcal{V} \equiv \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}$. Find $u \in \mathcal{V}$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathcal{V},$$

where

$$a(w, v) = \int_\Omega \nabla v \cdot \nabla w dx \quad \forall w, v \in \mathcal{V},$$
$$\ell(v) = \int_\Omega v f dx + \int_{\Gamma_N} v g ds \quad \forall v \in \mathcal{V}$$

for $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_N)$. The space $\mathcal{V}$ is endowed with the standard $H^1(\Omega)$ inner product and norm; i.e., $(\cdot,\cdot)_{\mathcal{V}} \equiv (\cdot,\cdot)_{H^1(\Omega)}$ and $\|\cdot\|_{\mathcal{V}} \equiv \|\cdot\|_{H^1(\Omega)}$.

We first show that the bilinear form is continuous. For all $w, v \in \mathcal{V}$,

$$|a(w, v)| = |\int_\Omega \nabla v \cdot \nabla w dx| \leq \|\nabla v\|_{L^2(\Omega)} \|\nabla w\|_{L^2(\Omega)} = |v|_{H^1(\Omega)} |w|_{H^1(\Omega)} \leq \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)};$$

here, the first inequality follows from Cauchy-Schwarz, and the last inequality follows from $|v|^2_{H^1(\Omega)} \leq |v|^2_{H^1(\Omega)} + \|v\|^2_{L^2(\Omega)} \equiv \|v\|^2_{H^1(\Omega)}$. Hence,

$$\gamma = \sup_{w \in \mathcal{V}} \sup_{v \in \mathcal{V}} \frac{|a(w, v)|}{\|w\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}} \leq 1 < \infty,$$

and the bilinear form is continuous.

We next show that the bilinear form is coercive. For all $v \in \mathcal{V}$,

$$\|v\|^2_{H^1(\Omega)} = |v|^2_{H^1(\Omega)} + \|v\|^2_{L^2(\Omega)} \leq |v|^2_{H^1(\Omega)} + C_{\rm PF} |v|^2_{H^1(\Omega)} = (1 + C_{\rm PF}) a(v, v),$$

where $C_{\rm PF} < \infty$ is the constant associated with the (corollary of the) Poincaré-Friedrichs inequality, Corollary 2.36. It follows that

$$\frac{a(v, v)}{\|v\|^2_{H^1(\Omega)}} \geq \frac{1}{1 + C_{\rm PF}} \quad \forall v \in \mathcal{V}.$$

Hence,

$$\alpha = \inf_{v \in \mathcal{V}} \frac{a(v,v)}{\|v\|^2_{H^1(\Omega)}} \geq \frac{1}{1 + C_{\rm PF}} > 0,$$

and the bilinear form is coercive.

We now show that the volume source term of the linear form is continuous. For all $v \in \mathcal{V}$,

$$|\int_\Omega v f dx| \leq \|v\|_{L^2(\Omega)} \|f\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)}; \tag{2.12}$$

here, the first inequality follows from Cauchy-Schwarz, and the second inequality follows from $\|v\|_{L^2(\Omega)} \leq (\|v\|^2_{L^2(\Omega)} + |v|^2_{H^1(\Omega)})^{1/2} = \|v\|_{H^1(\Omega)}$, $\forall v \in \mathcal{V}$. We then show that the Neumann boundary term of the linear form is continuous: $\forall v \in \mathcal{V}$,

$$|\int_{\Gamma_N} v g ds| \leq \|v\|_{L^2(\Gamma_N)} \|g\|_{L^2(\Gamma_N)} \leq C_{\rm tr} \|g\|_{L^2(\Gamma_N)} \|v\|_{H^1(\Omega)}; \tag{2.13}$$

here, the first inequality follows from Cauchy-Schwarz, and the second inequality follows from the trace inequality, Proposition 2.39. The combination of (2.12) and (2.13) yields, $\forall v \in \mathcal{V}$,

$$|\ell(v)| \leq |\int_\Omega v f dx| + |\int_{\Gamma_N} v g ds| \leq (\|f\|_{L^2(\Omega)} + C_{\rm tr} \|g\|_{L^2(\Gamma_N)}) \|v\|_{H^1(\Omega)}.$$

Hence,

$$\|\ell\|_{(H^1(\Omega))'} = \sup_{v \in \mathcal{V}} \frac{|\ell(v)|}{\|v\|_{H^1(\Omega)}} \leq \|f\|_{L^2(\Omega)} + C_{\rm tr} \|g\|_{L^2(\Gamma_N)} < \infty,$$

and ℓ is continuous.

We confirm that all conditions of the Lax-Milgram theorem are satisfied; hence the Poisson's problem has a unique solution, and the problem is well-posed.

2.12 Minimization formulation

We now obtain the minimization formulation for boundary value problems under two assumptions on the bilinear form $a(\cdot, \cdot)$:

1. coercivity: $\exists \alpha > 0$ such that $a(v,v) \geq \alpha \|v\|^2_{\mathcal{V}}$, $\forall v \in \mathcal{V}$.
2. symmetry: $a(w,v) = a(v,w)$, $\forall w, v, \in \mathcal{V}$.

Our variational formulation is as follows: find $u \in \mathcal{V}^E$ such that

$$a(u,v) = \ell(v) \quad \forall v \in \mathcal{V}, \tag{2.14}$$

where $\mathcal{V}^E \equiv \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = u^B\}$ for some boundary function u^B, $\mathcal{V} \equiv \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}$, and a continuous linear form $\ell : \mathcal{V} \to \mathbb{R}$. We then introduce an energy functional, $J : H^1(\Omega) \to \mathbb{R}$, given by

$$J(w) \equiv \frac{1}{2} a(w,w) - \ell(w) \quad \forall w \in H^1(\Omega). \tag{2.15}$$

We then have the following minimization formulation: find $u \in \mathcal{V}^E$ such that

$$u = \arg\min_{w \in \mathcal{V}^E} J(w). \tag{2.16}$$

This is a minimization formulation for a general symmetric, coercive problem.

We can readily show that $u \in \mathcal{V}$ is the solution to the minimization problem (2.16) if and only if it is the solution to the variational problem (2.14). Suppose $u \in \mathcal{V}^E$ is the solution to the variational problem (2.14). Let $w = u + v$ for some $v \in \mathcal{V}$. (Note that $w \in \mathcal{V}^E$.) We then observe, $\forall v \in \mathcal{V}$,

$$\begin{aligned} J(w) = J(u+v) &= \frac{1}{2}a(u+v,u+v) - \ell(u+v) \\ &= \underbrace{\frac{1}{2}a(u,u) - \ell(u)}_{J(u)} + \underbrace{\frac{1}{2}a(u,v) + \frac{1}{2}a(v,u)}_{=a(u,v) \text{ by symmetry}} - \ell(v) + \frac{1}{2}a(v,v) \\ &= J(u) + \underbrace{a(u,v) - \ell(v)}_{=0 \text{ since } u \text{ solves (2.14)}} + \underbrace{\frac{1}{2}a(v,v)}_{\substack{>0 \text{ for } v \neq 0 \\ \text{by coercivity}}} > J(u) \quad \forall v \neq 0. \end{aligned}$$

Hence, $J(w) > J(u)$ for all $w \neq u$, and u is the minimizer of J (2.15). Note that this proof relies on the fact that the bilinear form is symmetric and coercive.

Conversely, suppose $u \in \mathcal{V}^E$ is the solution to the minimization problem (2.16). Because the energy functional is quadratic in the argument, the minimizer u must satisfy the stationarity condition

$$J'(u;v) \equiv \lim_{\epsilon \to 0} \frac{1}{\epsilon}(J(u+\epsilon v) - J(u)) = 0 \quad \forall v \in \mathcal{V};$$

the Fréchet derivative (i.e., directional derivative) about u in any direction v should be 0. The Fréchet derivative $J'(u;v)$ is given by

$$\begin{aligned} J'(u;v) &\equiv \lim_{\epsilon \to 0} \frac{1}{\epsilon}(J(u+\epsilon v) - J(u)) = \lim_{\epsilon \to 0} \frac{1}{\epsilon}(J(u) + a(u,\epsilon v) - \ell(\epsilon v) + \frac{1}{2}a(\epsilon v, \epsilon v) - J(u)) \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon}(\epsilon a(u,v) - \epsilon \ell(v) + \frac{1}{2}\epsilon^2 a(v,v)) = \lim_{\epsilon \to 0}(a(u,v) - \ell(v) + \frac{1}{2}\epsilon a(v,v)) = a(u,v) - \ell(v) \end{aligned}$$

Hence, for $u \in \mathcal{V}^E$ to be the minimizer, it must satisfy

$$J'(u;v) = a(u,v) - \ell(v) = 0, \forall v \in \mathcal{V},$$

which is precisely the variational statement (2.14).

2.13 Summary

We summarize key points of this lecture:

1. A Hilbert space is a complete inner-product space; a Banach space is a complete normed space.

2. The Lebesgue space $L^2(\Omega)$ consists of functions that are square integrable (in the Lebesgue sense).

3. The Sobolev space $H^k(\Omega)$ consists of functions whose weak derivatives of up to and including order k are square integrable (in the Lebesgue sense).

4. Poisson's equation can be cast in the strong, minimization, or variational (or weak) form.

5. A Dirichlet boundary condition is an essential boundary condition that is imposed strongly by the choice of the space. Neumann and Robin boundary conditions are natural boundary conditions that are imposed weakly by the variational form.

6. Nonhomogeneous Dirichlet boundary conditions are imposed strongly by an affine (and not linear) trial space.

7. The Lax-Milgram theorem shows the existence *and* uniqueness of a solution for a weak formulation with a coercive and continuous bilinear from and a continuous linear form.

8. The verification of coercivity and continuity of a bilinear form often relies on a Poincaré-Friedrichs-type inequality and trace inequality, respectively.

9. If a variational formulation is given by a symmetric, coercive bilinear form, then it has a minimization formulation.

2.14 Appendix. Lax-Milgram: violation of the assumptions

Positive-definite but non-coercive $a(\cdot,\cdot)$. One of the required conditions of the Lax-Milgram theorem is coercivity: $\exists \alpha > 0$ such that $a(v,v) \geq \alpha \|v\|_{\mathcal{V}}^2 \ \forall v \in \mathcal{V}$. Note that this condition is stronger than the positive definiteness condition: $a(v,v) > 0 \ \forall v \neq 0$; in other words, coercivity implies positive definiteness but not the converse. We illustrate that the positive definiteness is insufficient to guarantee the existence of a solution using a concrete example.

We introduce a domain $\Omega \equiv (-1,1) \subset \mathbb{R}^1$, an associated Hilbert space $\mathcal{V} \equiv H^1(\Omega)$, and a bilinear form

$$a(w,v) \equiv \int_\Omega vw dx \quad \forall w, v \in \mathcal{V}.$$

We readily verify that the bilinear form is continuous with a continuity constant $\gamma = 1$. We also observe that the bilinear form is positive definite: $a(v,v) = \|v\|_{L^2(\Omega)}^2 > 0 \ \forall v \neq 0$. However, the form is not $\mathcal{V}$-coercive. To see this, consider a sequence of functions $v_n(x) = \sin(n\pi x)$. We observe that $\|v_n\|_{L^2(\Omega)}^2 = \int_\Omega v_n^2 dx = 1$ and $|v_n|_{H^1(\Omega)}^2 = \int_\Omega \left(\frac{dv_n}{dx}\right)^2 dx = n^2\pi^2$. It hence follows that

$$\alpha = \inf_{w\in\mathcal{V}} \frac{a(w,w)}{\|w\|_{H^1(\Omega)}} \leq \inf_{w\in\mathcal{V}} \frac{\|w\|_{L^2(\Omega)}}{|w|_{H^1(\Omega)}} \leq \inf_{n\in\mathbb{Z}_{>0}} \frac{\|v_n\|_{L^2(\Omega)}}{|v_n|_{H^1(\Omega)}} = \inf_{n\in\mathbb{Z}_{>0}} \frac{1}{n^2\pi^2} = 0.$$

Hence, the bilinear form is not $\mathcal{V}$-coercive (even though it is positive definite).

We now show that if the bilinear form is not $\mathcal{V}$-coercive, then the solution may not exist in $\mathcal{V}$. To this end, we introduce a linear form

$$\ell(v) \equiv \int_\Omega vf dx \quad \forall v \in \mathcal{V},$$

where

$$f(x) = \begin{cases} -1, & x \leq 0, \\ 1, & x > 0. \end{cases}$$

Note that $\|\ell\|_{\mathcal{V}'} \leq \|f\|_{L^2(\Omega)} = \sqrt{2}$, and hence ℓ is continuous. We then consider the following weak problem: find $u \in L^2(\Omega)$ such that

$$a(u,v) = \ell(v) \quad \forall v \in H^1(\Omega). \tag{2.17}$$

More explicitly, we seek $u \in L^2(\Omega)$ such that

$$\int_\Omega vu dx = \int_\Omega vf dx \quad \forall v \in H^1(\Omega).$$

We readily observe that a solution to this problem is $u = f$, which is in $L^2(\Omega)$ but not in $\mathcal{V} \equiv H^1(\Omega)$.

We now wish to show that $u = f$ is the unique solution to the problem, and in particular there is no solution $z \in H^1(\Omega)$ that satisfies $a(z,v) = \ell(v) \ \forall v \in H^1(\Omega)$. If both $u = f$ and z satisfies the weak statement (2.17), we observe that $a(u-z,v) = 0 \ \forall v \in H^1(\Omega)$. We now consider a sequence $v_n \in H^1(\Omega)$ such that $\|(u-z) - v_n\|_{L^2(\Omega)} \leq 1/n$; because $u - z$ is in $L^2(\Omega)$ and $H^1(\Omega)$ is dense in $L^2(\Omega)$, such a sequence exists. Since $v_n \in H^1(\Omega)$, it follows that

$$0 = a(u-z,v_n) = a(u-z,u-z) - a(u-z,(u-z)-v_n) = \|u-z\|_{L^2(\Omega)}^2 - (u-z,(u-z)-v_n)_{L^2(\Omega)}.$$

We rearrange the expression and invoke the Cauchy-Schwarz inequality to obtain

$$\|u-z\|^2_{L^2(\Omega)} = (u-z,(u-z)-v_n)_{L^2(\Omega)} \le \|u-z\|_{L^2(\Omega)}\|(u-z)-v_n\|_{L^2(\Omega)},$$

which implies

$$\|u-z\|_{L^2(\Omega)} \le \|(u-z)-v_n\|_{L^2(\Omega)} \le 1/n \to 0 \quad \text{as } n\to\infty.$$

Hence, we find that $u = f \in L^2(\Omega)$ (but not in $H^1(\Omega)$) is the unique solution to (2.17). Hence there does not exist a solution $u \in \mathcal{V} \equiv H^1(\Omega)$ such that $a(u,v) = \ell(v)\ \forall v \in \mathcal{V}$.

Non-continuous ℓ. We now consider a different case: we suppose that the bilinear form $a : \mathcal{V}\times\mathcal{V}\to\mathbb{R}$ is $\mathcal{V}$-coercive and $\mathcal{V}$-continuous, but $\ell \in \mathcal{V}'$ is not continuous. In other words, there exists a sequence of $v_n \in \mathcal{V}$ such that

$$\frac{|\ell(v_n)|}{\|v_n\|_{\mathcal{V}}} \to \infty \quad \text{as} \quad n\to\infty.$$

We now assume the solution such that $a(u,v) = \ell(v)\ \forall v\in\mathcal{V}$ exists. It then follows that

$$\frac{|\ell(v_n)|}{\|v_n\|_{\mathcal{V}}} = \frac{|a(u,v_n)|}{\|v_n\|_{\mathcal{V}}} \le \gamma\|u\|_{\mathcal{V}}.$$

Since $|\ell(v_n)|/\|v_n\|_{\mathcal{V}} \to \infty$ and $\gamma < \infty$, we conclude that $\|u\|_{\mathcal{V}}$ is not bounded and does not belong to $\mathcal{V}$.

Lecture 3

Finite element method: formulation

3.1 Introduction

In this lecture, we develop finite element approximations of variational problems. Specifically, we introduce a triangulation of the domain Ω, construct an approximation space $\mathcal{V}_h \subset \mathcal{V}$, formulate a discrete problem, and then discuss the well-posedness of the finite element problem.

3.2 Triangulation

We first introduce a *triangulation* of $\Omega \subset \mathbb{R}^d$. A triangulation

$$\mathcal{T}_h \equiv \{K_i\}_{i=1}^{n_e}$$

is a set of non-overlapping elements $K_1, \dots, K_{n_e}$ such that the union of the closure of the elements covers the domain:

$$\begin{aligned} K_i \cap K_j &= \emptyset, \quad i \neq j \\ \cup_{i=1}^{n_e} \overline{K}_i &= \overline{\Omega}. \end{aligned}$$

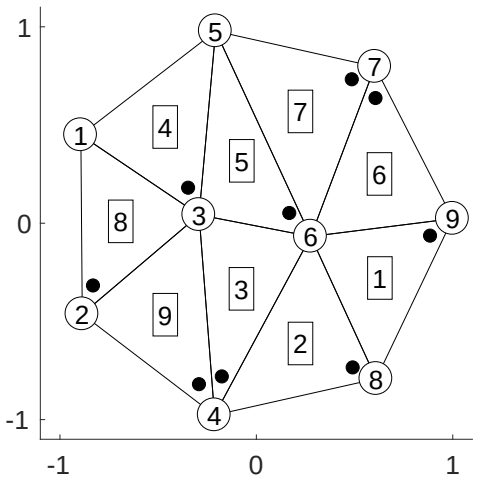
(We consider the closure of the elements because we consider each element to be open.) An example of a triangulation is shown in Figure 3.1. The triangulation comprises $n_e = 9$ triangular elements $\{K_i\}_{i=1}^{n_e}$, which are delineated by $n_v = 9$ vertices $\{z_i\}_{i=1}^{n_v}$. For each element K_i, we define the *diameter*

$$h_{K_i} \equiv \operatorname{diam}(K_i).$$

The diameter of K_i is the supremum of the distances between pairs of points in K^i,

$$\operatorname{diam}(K_i) = \sup_{x,y \in K_i} \|x - y\|_2;$$

the diameter for a triangle K_i is the length of the longest edge. (We take the supremum because each element is open.) The subscript h of the triangulation $\mathcal{T}_h$ signifies the maximum diameter of



(a) mesh

Figure 3.1: Triangulation.

the elements in the triangulation,

$$h \equiv \max_{i=1,\ldots,n_e} h_{K_i}.$$

Intuitively (and as we will see more formally), the finite element spaces associated with a sequence of triangulations get richer as h decreases. In general, a triangulation comprises line segments in one dimension, triangles or quadrilaterals in two dimension, and tetrahedrons or hexahedrons in three dimensions.

While mathematically a triangulation is simply a collection of non-overlapping elements that cover the domain, we need a convenient means to represent the triangulation on a computer. One approach is to store tables of *node coordinates* and *element-node connectivities*. Tables 3.1(a) and 3.1(b) are respectively the coordinate and connectivity tables associated with the triangulation shown in Figure 3.1. The connectivity table indicates that, for instance, the element K_5 is delineated by the nodes z_6, z_5, and z_3; the coordinate table then indicates that the coordinates of these three nodes are $z_6 = (0.28, -0.07)$, $z_5 = (-0.21, 0.98)$, and $z_3 = (-0.29, 0.04)$. By convention, we order the nodes of the triangles in the counterclockwise manner. The coordinate and connectivity tables together provide a complete geometric description of all elements. Note that, in Figure 3.1, for each triangle, we indicate the first of the three nodes that delineate the triangle by a dot (•); with this convention we have identical information presented in Figure 3.1 in a visual form and Tables 3.1(a) and 3.1(b) in an array form.

The task of generating a triangulation for a given domain is called *mesh generation* and a software that carries out the task is called a *mesh generator* or *mesher*. Mesh generation is a non-trivial task. In fact, the development of algorithms that can robustly and automatically generate high-quality triangulation for complex geometries in three dimensions is an area of ongoing research. Nevertheless, because mesh generation is essential for any finite element discretization, there are many commercial and open-source meshers. Here we name a few user-friendly, open-source meshers:

- `triangle`. A robust two-dimensional mesher written in C that generates meshes with a guaranteed quality certificate in terms of the minimal angle.

Table 3.1: Node coordinate and connectivity table for mesh shown in Figure 3.1.

(a) coordinates

node	x_1	x_2
1	−0.89	0.45
2	−0.89	−0.46
3	−0.29	0.04
4	−0.21	−0.98
5	−0.21	0.98
6	0.28	−0.07
7	0.60	0.80
8	0.61	−0.79
9	1.00	0.02

(b) connectivity

element	node 1	node 2	node 3
1	9	6	8
2	8	6	4
3	4	6	3
4	3	5	1
5	6	5	3
6	7	6	9
7	7	5	6
8	2	3	1
9	4	3	2

- `tetgen`. A popular three-dimensional mesher written in C.
- `distmesh`. A user-friendly mesher written in MATLAB for implicit domain geometries represented by level sets.

The mesh shown in Figure 3.1 was in fact generated by `distmesh`. We will extensively use `distmesh` to generate meshes in this course as it is implemented in MATLAB and is easy to use.

3.3 Approximation spaces

We now introduce *approximation spaces* (or *finite element spaces*) for $\mathcal{V}$. An approximation space is a finite-dimensional subspace space of $\mathcal{V}$ with which we can approximate functions in $\mathcal{V}$. For concreteness, we consider a piecewise linear approximation space for $\mathcal{V} \equiv H^1(\Omega)$ associated with the triangulation $\mathcal{T}_h$,

$$\mathcal{V}_h \equiv \{v \in \mathcal{V} \equiv H^1(\Omega) \mid v|_K \in \mathbb{P}^1(K),\ \forall K \in \mathcal{T}_h\}; \tag{3.1}$$

here $\mathbb{P}^1(K)$ is the space of linear polynomials over K. We note the two requirements: $v \in \mathcal{V}_h$ must belong to $\mathcal{V} \equiv H^1(\Omega)$; v restricted to any element $K \in \mathcal{T}_h$, $v|_K$, must be a linear polynomial. Figure 3.2 shows an example of a function in a linear ($\mathbb{P}^1$) finite element space, associated with the mesh shown in Figure 3.1.

We note that the condition $\mathcal{V}_h \subset H^1(\Omega)$ means that the weak derivative of functions must be square integrable (in the Lebesgue sense); for piecewise polynomials, the condition is satisfied if and only if the function is continuous. To see this, we observe the following. If a function is continuous and piecewise polynomial, the weak first derivative is a (potentially discontinuous) piecewise polynomial and hence is square integrable; the function hence is in $H^1(\Omega)$. If a function is not continuous across element interfaces, then the weak first derivative generates delta distributions at the interfaces and hence is not in $L^2(\Omega)$; recall for instance a concrete example for a Heaviside-like function in Section 2.3. Hence, for a piecewise polynomial function, the continuity is a necessary and sufficient condition for the function to be in $H^1(\Omega)$.

We now need a convenient means to describe functions in $\mathcal{V}_h$ given by (3.1), such as the one shown in Figure 3.2. Specifically, we need to pick *global degrees of freedom* with which we can

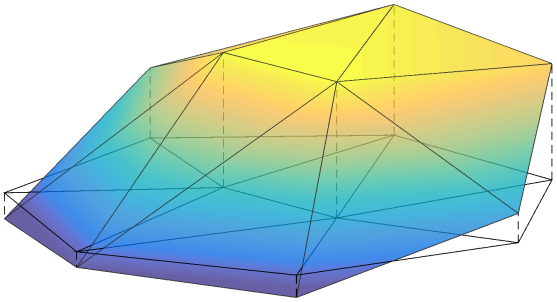


Figure 3.2: A function in a linear finite element space.

uniquely associate any function in $\mathcal{V}_h$ to a set of real numbers. To this end, we introduce a *basis* for the linear space $\mathcal{V}_h$. We recall that a set of functions $\{\phi_i\}_{i=1}^n$ is a basis for $\mathcal{V}_h$ if the set (i) spans $\mathcal{V}_h$ and (ii) is linearly independent. The first requirement implies that any $w \in \mathcal{V}_h$ can be expressed as a linear combination of $\{\phi_i\}_{i=1}^n$. The second requirement implies that the coefficients associated with the representation of $w \in \mathcal{V}_h$ in terms of $\{\phi_i\}_{i=1}^n$ is unique. In other words, if $\{\phi_i\}_{i=1}^n$ is a basis for $\mathcal{V}_h$, then for any $w \in \mathcal{V}_h$ there exists a unique $\hat{w} \in \mathbb{R}^n$ such that

$$w = \sum_{j=1}^{n} \hat{w}_j \phi_j \tag{3.2}$$

for $n = \dim(\mathcal{V}_h)$. Given a basis $\{\phi_i\}_{i=1}^n$ for $\mathcal{V}_h$, the relationship (3.2) is an isomorphism (i.e., a bijective map) from $\mathbb{R}^n$ to $\mathcal{V}_h$. A function that belongs to a basis $\{\phi_i\}_{i=1}^n$ is called a *basis function* or a *shape function*.

While the choice of a basis is not unique, one convenient choice is a *Lagrange basis* or *nodal basis*. A nodal basis $\{\phi_j\}_{j=1}^n$ comprises functions that take on the value of 1 at the associated node and 0 at all other nodes:

$$\phi_j(z_i) = \delta_{ij}, \tag{3.3}$$

where z_i is the i-th node of the triangulation, and δ_{ij} is the Kronecker delta so that $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. Figure 3.3 shows an example of a basis function, ϕ_3, for the linear finite element space (3.1) associated with the mesh shown in Figure 3.1. For $\mathcal{V}_h$ defined by (3.1), there are nine nodal basis functions, one associated with each node. We also observe that the set of the nine functions indeed forms a basis: the set is linearly independent and spans the space. The set is linearly independent because $\sum_{j=1}^n \hat{w}_j \phi_j = 0$ implies $\sum_{j=1}^n \hat{w}_j \phi_j(z_i) = 0$, $\forall i = 1, \ldots, n$, which in turn implies $\sum_{j=1}^n \hat{w}_j \delta_{ij} = \hat{w}_i = 0$, $i = 1, \ldots, n$. The set spans the space because the piecewise linear polynomial space is nine dimensional and the linearly independent set contains nine functions.

The nodal basis, unlike many other bases, provides a convenient interpretation in the physical space. Specifically, for $w \in \mathcal{V}_h$, we have a (unique) representation

$$w = \sum_{j=1}^{n} \hat{w}_j \phi_j = \sum_{j=1}^{n} w(z_j) \phi_j. \tag{3.4}$$

We note that the coefficient $\hat{w}_j$ must be equal to $w(z_j)$, because $w(z_i) = \sum_{j=1}^n \hat{w}_j \phi_j(z_i) = \sum_{j=1}^n \hat{w}_j \delta_{ij} = \hat{w}_i$, $i = 1, \ldots, n$; here, the second equality follows from the Lagrange interpolation property (3.3). In words, $\hat{w}_j = w(z_j)$, the value of the function $w \in \mathcal{V}_h$ evaluated at the

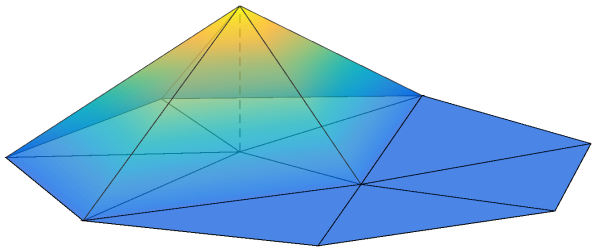


Figure 3.3: Nodal basis ϕ_3 for the linear finite element space $\mathcal{V}_h$ defined by (3.1).

associated node x_j. This interpretation of nodal basis functions also allows us to readily confirm that the nodal basis is indeed a basis: for any $w \in \mathcal{V}_h$, there exists a unique $\hat{w} \in \mathbb{R}^n$ such that $w = \sum_{j=1}^n \hat{w}_j \phi_j$.

We note that approximation spaces of the form (3.1) can be refined to yield a sequence of approximation spaces. A space $\mathcal{V}_{h'}$ is said to be a *refinement* of a space $\mathcal{V}_h$ if

$$\mathcal{V}_h \subset \mathcal{V}_{h'};$$

i.e., every member of $\mathcal{V}_h$ is also a member of $\mathcal{V}_{h'}$. For a piecewise linear space, a refinement results from splitting some or all of elements in the triangulation. Through a successive refinement of elements, we can construct a sequence of approximation spaces

$$\mathcal{V}_{h_1} \subset \mathcal{V}_{h_2} \subset \cdots \subset \mathcal{V}_{h_n}$$

for $h_1 > h_2 > \cdots > h_n$. (We recall that the subscript h of $\mathcal{V}_h$ indicates the diameter of the largest element, $h \equiv \max_{K \in \mathcal{T}_h} \text{diam}(K)$.) The ability to refine, and hence construct a sequence of enriched spaces, is important. Intuitively, we might relate this ability to arbitrary refine the approximation spaces with the ability to find an approximation that is arbitrary close to the exact solution. We will make this notion of convergence more formal in a later lecture.

3.4 Approximation spaces: essential boundary conditions

We recall from Lecture 2 that Dirichlet boundary conditions are treated as essential boundary conditions in the weak formulation of second-order elliptic PDEs. The essential boundary conditions are explicitly imposed through the choice of the space. For instance, given a mixed Poisson problem on $\Omega \subset \mathbb{R}^d$ with a homogeneous Dirichlet boundary $\Gamma_D \subset \partial\Omega$, the appropriate function space is

$$\mathcal{V} \equiv \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}.$$

We wish to construct an approximation space $\mathcal{V}_h \subset \mathcal{V}$.

To facilitate our discussion, we first introduce a piecewise linear approximation space for $H^1(\Omega)$ (without the essential boundary condition),

$$H^1_h(\Omega) \equiv \{v \in H^1(\Omega) \mid v|_K \in \mathbb{P}^1(K),\ \forall K \in \mathcal{T}_h\};$$

we note that the notation $H^1_h(\Omega)$ is *not* standard in literature, but we adhere to it as it is convenient. We then introduce a piecewise linear approximation for $\mathcal{V} \subset H^1(\Omega)$ with the essential boundary condition,

$$\mathcal{V}_h \equiv \{v \in \mathcal{V} \mid v|_K \in \mathbb{P}^1(K),\ \forall K \in \mathcal{T}_h\};$$

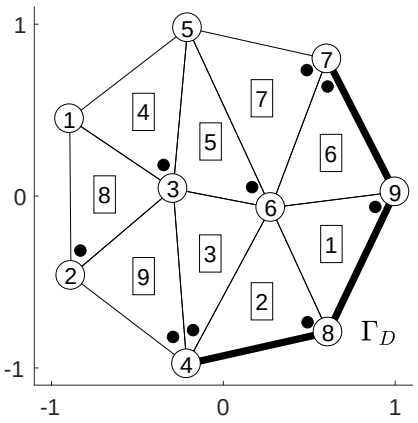


Figure 3.4: A triangulated domain with a Dirichlet boundary.

functions in the approximation space $\mathcal{V}_h$ must vanish on Γ_D so that $\mathcal{V}_h \subset \mathcal{V}$. The space $\mathcal{V}_h$ is a subspace of $\mathcal{V}$ because the functions restricted to element K is in $\mathbb{P}^1(K)$. The space $\mathcal{V}_h$ is also a subspace of $H^1_h(\Omega)$ because, while both spaces comprise piecewise linear functions, the functions in $\mathcal{V}_h$ must vanish on the boundary Γ_D.

For a piecewise linear function $v \in H^1_h(\Omega)$, a condition equivalent to $v|_{\Gamma_D} = 0$ is

$$v_h(z_j) = 0 \quad \text{for all nodes } z_j \text{ on } \bar{\Gamma}_D,$$

where $\bar{\Gamma}_D$ is the closure of the Dirichlet boundary. We consider the closure so that the nodes at the boundary (i.e., endpoints in $\mathbb{R}^2$) are included in the set. In other words,

$$\mathcal{V}_h = \{v \in H^1_h(\Omega) \mid v(z_j) = 0, \ \forall z_j \text{ on } \bar{\Gamma}_D\}.$$

For instance, in Figure 3.4, the piecewise linear function must vanish on the nodes z_4, z_7, z_8, and z_9.

For a nodal basis, the boundary condition can be explicitly imposed by eliminating the shape functions associated with the nodes on $\overline{\Gamma}_D$ from the approximation space. The dimension of the resulting approximation space for $\mathcal{V} \subset H^1(\Omega)$ is

$$n \equiv \dim(\mathcal{V}_h) = \dim(H^1_h(\Omega)) - (\text{number of nodes on } \bar{\Gamma}_D).$$

For instance, the dimension of the piecewise linear approximation space shown in Figure 3.4 is $n = 9 - 4 = 5$, where the active shape functions are associated with the nodes z_1, z_2, z_3, z_5, and z_6. Once the set of active shape functions of $H^1_h(\Omega)$ are identified, we can readily reassign them as a basis $\{\phi_j\}_{j=1}^n$ of $\mathcal{V}_h$. Without loss of generality, we also reassign the associated node numbers $\{z_j\}_{j=1}^n$. Then, as before, we can express any function in $w \in \mathcal{V}_h$ in terms of $\hat{w} \in \mathbb{R}^n$ as

$$w = \sum_{j=1}^{n} \hat{w}_j \phi_j = \sum_{j=1}^{n} w(z_j)\phi_j.$$

We now consider an nonhomogeneous Dirichlet boundary condition, say $u|_{\Gamma_D} = u^B$. We recall that the appropriate test space for the problem is $\mathcal{V} \equiv \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}$ and the trial

space is $\mathcal{V}^E \equiv u^E + \mathcal{V}$, where u^E is an arbitrary member of $H^1(\Omega)$ that satisfies the boundary condition $u^E|_{\Gamma_D} = u^B$. If the boundary function u^B is a piecewise polynomial that conforms to the triangulation, then it is possible to find $u_h^E \in H_h^1(\Omega)$ that satisfies the boundary condition exactly: $u_h^E|_{\Gamma_D} = u^B$. Otherwise, we have to choose $u_h^E \in H_h^1(\Omega)$ in the piecewise polynomial space that approximately satisfies the boundary condition: $u_h^E|_{\Gamma_D} \approx u^B$. In either case, a convenient choice (though not the only choice) is to simply choose any $u_h^E \in H_h^1(\Omega)$ such that

$$u_h^E(z_j) = u^B(z_j) \quad \text{for all nodes } z_j \text{ on } \bar{\Gamma}_D.$$

We can then express any function in $w \in \mathcal{V}_h^E \equiv u_h^E + \mathcal{V}_h$ in terms of $\hat{w} \in \mathbb{R}^n$ as

$$w = u_h^E + \sum_{j=1}^n \hat{w}_j \phi_j,$$

where $\{\phi_j\}_{j=1}^n$ is a nodal basis associated with nodes not on $\bar{\Gamma}_D$.

3.5 Galerkin method

We now consider a *Galerkin finite element approximation* of a boundary value problem. We first recall the weak formulation for the exact problem: find $u \in \mathcal{V}$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathcal{V}, \tag{3.5}$$

where $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is a coercive, continuous bilinear form and $\ell : \mathcal{V} \to \mathbb{R}$ is a continuous linear form. (In general, the bilinear form needs not be coercive; however, here we assume coercivity to prove some theoretical results using the tools introduced in the previous lecture.) We now seek an approximation to (3.5) in a finite-dimensional subspace $\mathcal{V}_h \subset \mathcal{V}$: find $u_h \in \mathcal{V}_h$ such that

$$a(u_h, v) = \ell(v) \quad \forall v \in \mathcal{V}_h. \tag{3.6}$$

This is the Galerkin finite element approximation of (3.5). In words, we obtain the finite element problem by simply restricting the test and trials spaces from $\mathcal{V}$ to $\mathcal{V}_h \subset \mathcal{V}$. Because the trial and test approximation spaces are the same, the method is referred to as a *Galerkin method*, or, more explicitly, *Bubnov-Galerkin method*. (If the test and trial approximation spaces are different, the method is referred to as a *Petrov-Galerkin method*.) The finite element problem (3.6) depends on the space $\mathcal{V}_h$ but is independent of the particular basis $\{\phi_i\}_{i=1}^n$ for $\mathcal{V}_h$. We will prove in Section 3.6 that the solution to (3.6) exists and is unique.

We now wish to recast the finite element problem (3.6) in linear algebraic from that is amenable to computer implementation. To this end, we represent the solution and test functions in terms of their basis coefficients, $u_h = \sum_{j=1}^n \hat{u}_{h,j}\phi_j$ and $v = \sum_{i=1}^n \hat{v}\phi_i$ to yield the following equivalent problem for the coefficients: find $\hat{u}_h \in \mathbb{R}^n$ such that

$$a(\sum_{j=1}^n \hat{u}_{h,j}\phi_j, \sum_{i=1}^n \hat{v}_i\phi_i) = \ell(\sum_{i=1}^n \hat{v}_i\phi_i) \quad \forall \hat{v} \in \mathbb{R}^n.$$

We then invoke the bilinearity of $a(\cdot, \cdot)$ and the linearity of $\ell(\cdot)$ to obtain

$$\sum_{i,j=1}^n \hat{v}_i a(\phi_j, \phi_i)\hat{u}_{h,j} = \sum_{i=1}^n \hat{v}_i \ell(\phi_i).$$

The problem can be more compactly expressed using the matrix-vector notation: find $\hat{u}_h \in \mathbb{R}^n$ such that

$$\hat{v}^T \hat{A}_h \hat{u}_h = \hat{v}^T \hat{f}_h \quad \forall \hat{v} \in \mathbb{R}^n, \tag{3.7}$$

where the *stiffness matrix* $\hat{A}_h \in \mathbb{R}^{n \times n}$ is given by

$$\hat{A}_{h,ij} \equiv a(\phi_j, \phi_i), \quad i, j = 1, \ldots, n,$$

and the *load vector* $\hat{f}_h \in \mathbb{R}^n$ is given by

$$\hat{f}_{h,i} \equiv \ell(\phi_i), \quad i = 1, \ldots, n.$$

In order for (3.7) to hold, each row of $A_h \hat{u}_h - f_h$ must be equal to zero; otherwise, we can find $\hat{v} \in \mathbb{R}^n$ that is finite only on that non-zero and hence (3.7) would not hold. We hence conclude that the statement (3.7) is equivalent to finding $\hat{u}_h \in \mathbb{R}^n$ that satisfies

$$\hat{A}_h \hat{u}_h = \hat{f}_h \quad (\text{in } \mathbb{R}^n). \tag{3.8}$$

We will prove in Section 3.6 that (3.8) has a unique solution.

3.6 Well-posedness of the Galerkin finite element formulation

The following proposition shows that the solution to (3.6) exists and is unique.

Proposition 3.1. Given an approximation space $\mathcal{V}_h \subset \mathcal{V}$, a continuous, coercive (but not necessarily symmetric) bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, and a continuous linear functional $\ell \in \mathcal{V}'$, there exists a unique $u_h \in \mathcal{V}_h$ such that

$$a(u_h, v) = \ell(v) \quad \forall v \in \mathcal{V}_h.$$

Proof. We will appeal to the Lax-Milgram theorem, Theorem 2.32. To this end, we need to demonstrate that (i) the bilinear form $a(\cdot, \cdot)$ is coercive in $\mathcal{V}_h$, (ii) the bilinear form is continuous $a(\cdot, \cdot)$ is continuous in $\mathcal{V}_h$, and (iii) the linear form $\ell(\cdot)$ is continuous in $\mathcal{V}_h$. All of these properties follow from the fact that $\mathcal{V}_h$ is a subspace of $\mathcal{V}$. The coercivity of $a(\cdot, \cdot)$ in $\mathcal{V}_h$ follows from

$$\alpha_h \equiv \inf_{v \in \mathcal{V}_h} \frac{a(v, v)}{\|v\|_{\mathcal{V}}^2} \geq \inf_{v \in \mathcal{V}} \frac{a(v, v)}{\|v\|_{\mathcal{V}}^2} \equiv \alpha > 0,$$

where the inequality follows from $\mathcal{V}_h \subset \mathcal{V}$. The coercivity constant associated with $\mathcal{V}_h$, α_h, is bounded from the below by the coercivity constant associated with $\mathcal{V}$, α, which itself is bounded from the below by 0. The continuity of $a(\cdot, \cdot)$ in $\mathcal{V}_h$ follows from

$$\gamma_h \equiv \sup_{w, v \in \mathcal{V}_h} \frac{|a(w, v)|}{\|w\|_{\mathcal{V}} \|v\|_{\mathcal{V}}} \leq \sup_{w, v \in \mathcal{V}} \frac{|a(w, v)|}{\|w\|_{\mathcal{V}} \|v\|_{\mathcal{V}}} \equiv \gamma < \infty,$$

where the inequality again follows from $\mathcal{V}_h \subset \mathcal{V}$. The continuity constant associated with $\mathcal{V}_h$, γ_h, is bounded from the above by the continuity constant associated with $\mathcal{V}$, γ, which itself is finite. Similarly, the continuity of $\ell(\cdot)$ in $\mathcal{V}_h$ follows from

$$\|\ell\|_{(\mathcal{V}_h)'} \equiv \sup_{v \in \mathcal{V}_h} \frac{|\ell(v)|}{\|v\|_{\mathcal{V}}} \leq \sup_{v \in \mathcal{V}} \frac{|\ell(v)|}{\|v\|_{\mathcal{V}}} \equiv \|\ell\|_{\mathcal{V}'} < \infty,$$

where the inequality again follows from $\mathcal{V}_h \subset \mathcal{V}$. Because the bilinear from is coercive and continuous in $\mathcal{V}_h$ and the linear from is continuous in $\mathcal{V}_h$, we conclude by the Lax-Milgram theorem that the solution exists and is unique. □

Proposition 3.1 shows the existence of a unique solution to the finite element problem (3.6) without appealing to any specific basis for $\mathcal{V}_h$. That is, the finite element solution $u_h \in \mathcal{V}_h$ depends only on the space $\mathcal{V}_h$ and is independent of the specific basis $\{\phi_i\}_{i=1}^n$ used to represent functions in $\mathcal{V}_h$.

For a problem that involves nonhomogeneous Dirichlet data, the well-posedness of the finite element problem can be proved using the same technique used for the exact (infinite-dimensional) problem in Section 2.9; we first reformulate it as a homogeneous Dirichlet problem with a modified linear form and then apply the Lax-Milgram theorem.

We now consider the well-posedness of the linear algebraic problem (3.8).

Proposition 3.2. Under the condition of Proposition 3.1, introduce a basis $\{\phi_i\}_{i=1}^n$ for $\mathcal{V}_h$. There exists a unique solution $\hat{u}_h \in \mathbb{R}^n$ to

$$\hat{A}_h \hat{u}_h = \hat{f}_h,$$

where $\hat{A}_{h,ij} \equiv a(\phi_j, \phi_i)$ and $\hat{f}_{h,i} \equiv \ell(\phi_i)$.

Proof. Proposition 3.1 shows the existence and uniqueness of the solution $\hat{u}_h \in \mathcal{V}_h$. Because $\{\phi_i\}_{i=1}^n$ is a basis for $\mathcal{V}_h$, there exists a unique coefficients $\hat{u}_h \in \mathbb{R}^n$ such that

$$u_h = \sum_{j=1}^n \hat{u}_{h,j} \phi_j.$$

The coefficients $\hat{u}_h \in \mathbb{R}^n$ is the unique solution of the linear system $\hat{A}_h \hat{u}_h = \hat{f}_h$. □

Alternatively, we can show that, if $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is coercive, then the matrix $\hat{A}_h \in \mathbb{R}^{n \times n}$ is *positive definite* and hence is non-singular, and a unique solution to $\hat{A}_h \hat{u}_h = \hat{f}_h$ exists.

Definition 3.3 (positive definite matrix)**.** A matrix $A \in \mathbb{R}^{n \times n}$, not necessarily symmetric, is positive definite if

$$x^T A x \geq 0 \quad \forall x \in \mathbb{R}^n,$$

and $x^T A x = 0$ if and only if $x = 0$.

Proposition 3.4. If the bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is coercive (but not necessarily symmetric), then the stiffness matrix $\hat{A}_h \in \mathbb{R}^{n \times n}$ associated with a basis $\{\phi_i\}_{i=1}^n$ of $\mathcal{V}_h$ is positive definite (PD).

Proof. We observe that $\forall \hat{v} \in \mathbb{R}^n$,

$$\hat{v}^T \hat{A}_h \hat{v} = \sum_{i,j=1}^n \hat{v}_i a(\phi_j, \phi_i) \hat{v}_j = a(\sum_{j=1}^n \hat{v}_j \phi_j, \sum_{i=1}^n \hat{v}_i \phi_i) \geq \alpha \| \sum_{i=1}^n \hat{v}_i \phi_i \|_{\mathcal{V}}^2 \geq 0,$$

where the first inequality follows from the coercivity of $a(\cdot, \cdot)$ and the last equality follows from the positive definiteness of the norm $\| \cdot \|_{\mathcal{V}}$. Moreover, because $\| \cdot \|_{\mathcal{V}}$ is a norm, $\| \sum_{i=1}^n \hat{v}_i \phi_i \|_{\mathcal{V}} = 0$ if and only if $\sum_{i=1}^n \hat{v}_i \phi_i = 0$. In addition, because $\{\phi_i\}_{i=1}^n$ is a basis and in particular linearly independent, $\sum_{i=1}^n \hat{v}_i \phi_i = 0$ if and only if $\hat{v}_i = 0$, $\forall i = 1, \ldots, n$. It follows that

$$\hat{v}^T \hat{A}_h \hat{v} = 0 \quad \text{if and only if} \quad \hat{v} = 0.$$

Hence, the matrix $\hat{A}_h \in \mathbb{R}^{n \times n}$ is PD. □

Proposition 3.5. A positive definite (not necessarily symmetric) matrix $A \in \mathbb{R}^{n\times n}$ is non-singular.

Proof. The proof is by contraction. Suppose the matrix A is singular. Then, $\ker(A) \neq 0$, and there exists $x \neq 0$ such that $Ax = 0$. This in turn implies that $x^T Ax = 0$ for the vector $x \neq 0$. But if A is PD, then $x^T Ax = 0$ if and only if $x = 0$. We hence have a contraction; a PD matrix must have $\ker(A) = 0$ and hence is non-singular. □

Remark 3.6. Because the matrix $\hat{A}_h \in \mathbb{R}^{n\times n}$ is PD, the matrix is non-singular, and the linear system $\hat{A}_h \hat{u}_h = \hat{f}_h$ has a unique solution.

If the bilinear form is not only coercive but also symmetric, then we can also show that the matrix $\hat{A}_h \in \mathbb{R}^{n\times n}$ is symmetric positive definite (SPD).

Definition 3.7 (symmetric positive definite matrix)**.** A matrix $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite (SPD) if A is

(i) symmetric: $A^T = A$

(ii) positive definite: $x^T Ax \geq 0 \ \forall x \in \mathbb{R}^n$, and $x^T Ax = 0$ if and only if $x = 0$.

Proposition 3.8. If the bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is symmetric and coercive, then the stiffness matrix $\hat{A}_h \in \mathbb{R}^{n\times n}$ is symmetric positive definite (SPD).

Proof. The symmetry of $\hat{A}_h$ follows directly from the symmetry of the bilinear form:

$$\hat{A}_{h,ij} = a(\phi_j, \phi_i) = a(\phi_i, \phi_j) = \hat{A}_{h,ji} \quad \forall i, j = 1, \ldots, n.$$

We have already shown in Proposition 3.4 that the coercivity of $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ implies the positive definiteness of $\hat{A}_h$. □

Before we conclude a section, we make an important remark. A finite element solution $u_h \in \mathcal{V}_h$ is a *field* that approximates the solution $u \in \mathcal{V}$. Because u_h is a field, we can evaluate $u_h(x)$ at any point $x \in \Omega$. This is fundamentally different from a finite difference method, which approximates the solution at a discrete set of points, and the solution elsewhere is in general not approximated. (Of course, we could in practice interpolate the finite difference solution, but that involves an additional approximation.) While the coefficients $\hat{u}_h \in \mathbb{R}^n$ is associated with the nodal value of the solution for nodal bases, this is a rather special interpretation for a nodal bases. The finite element solution field $u_h \in \mathcal{V}_h$ only depends on the space $\mathcal{V}_h$ and not the particular basis used to represent the functions in $\mathcal{V}_h$.

3.7 Minimization formulation

For a variational problem with a symmetric, coercive bilinear form, we can also formulate the finite element problem using a minimization formulation. We recall that the energy functional associated with the minimization problem is $J : \mathcal{V} \to \mathbb{R}$ such that

$$J(w) \equiv \frac{1}{2} a(w, w) - \ell(w),$$

where $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is a symmetric, coercive bilinear form, and $\ell : \mathcal{V} \to \mathbb{R}$ is a linear form. The solution is given by $u \in \mathcal{V}$ such that

$$u = \arg\min_{w \in \mathcal{V}} J(w).$$

We recall that the minimization formulation and the variational formulation are equivalent.

The finite element approximation is given by the minimization problem over a subspace $\mathcal{V}_h \subset \mathcal{V}$. To this end, we introduce a basis $\{\phi_i\}_{i=1}^n$ for $\mathcal{V}_h$ and evaluate the energy functional for an arbitrary $w \equiv \sum_{j=1}^n \hat{w}_j \phi_j$ in the space:

$$J(w) = J(\sum_{j=1}^n \hat{w}_j \phi_j) = \frac{1}{2} a(\sum_{j=1}^n \hat{w}_j \phi_j, \sum_{i=1}^n \hat{w}_i \phi_i) - \ell(\sum_{i=1}^n \hat{w}_i \phi_i) = \frac{1}{2} \sum_{i,j=1}^n \hat{w}_i \underbrace{a(\phi_j, \phi_i)}_{\hat{A}_{h,ij}} \hat{w}_j - \sum_{i=1}^n \hat{w}_i \underbrace{\ell(\phi_i)}_{\hat{f}_{h,i}};$$

here identify the stiffness matrix $\hat{A}_h$ and load vector $\hat{f}_h$ in the expression. We can then redefine an energy functional in terms of the coefficients $\hat{w} \in \mathbb{R}^n$ such that

$$\hat{J}(\hat{w}) \equiv J(\sum_{j=1}^n \hat{w}_j \phi_j) = \frac{1}{2} \hat{w}^T \hat{A}_h \hat{w} - \hat{w}^T \hat{f}_h.$$

We then seek the coefficients $\hat{u}_h \in \mathbb{R}^n$ such that

$$\hat{u}_h = \arg\min_{\hat{w} \in \mathbb{R}^n} \hat{J}(\hat{w}).$$

The sufficient condition for $\hat{u}_h$ to be the minimizer is that (i) the gradient is zero, $\nabla \hat{J}(\hat{u}_h) = 0$, and (ii) the Hessian of $J(\hat{u}_h)$ is SPD. The condition (i) is equivalent to

$$\nabla \hat{J}(\hat{u}_h) = \hat{A}_h \hat{u}_h - \hat{f}_h = 0 \quad (\text{in } \mathbb{R}^n),$$

which is the same as the Galerkin finite element statement for the coefficients, (3.8). The condition (ii) is also satisfied because the Hessian of $J(\hat{u}_h)$ is $\hat{A}_h$, which we have proven is SPD for a symmetric, coercive bilinear form in Proposition (3.8). Hence, the finite element solution $u_h \in \mathcal{V}_h$, and the associated coefficients $\hat{u}_h \in \mathbb{R}^n$, can also be obtained using the minimization formulation.

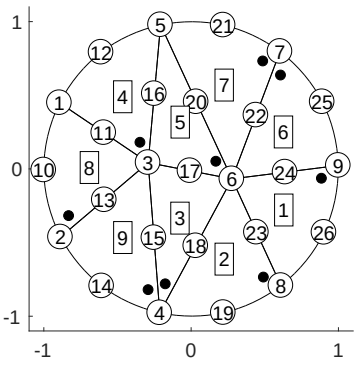
3.8 Generalization: higher-order and spectral methods

In this lecture, we constructed the Galerkin approximation based on an approximation space $\mathcal{V}_h$ of piecewise linear ($\mathbb{P}^1$) polynomials; however, the Galerkin method — which approximates the solution $u \in \mathcal{V}$ in a subspace $\mathcal{V}_h$ — is in fact a general procedure that works with any approximation space $\mathcal{V}_h \subset \mathcal{V}$.

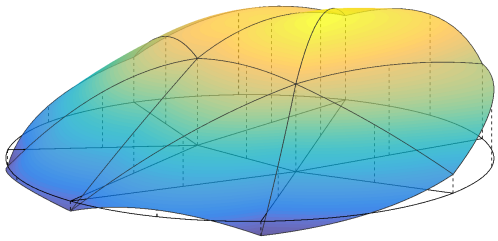
For instance, we could consider a domain with a curved boundary and an associated triangulation as shown in Figure 3.5(a), and then considered a space of piecewise quadratic ($\mathbb{P}^2$) polynomials,

$$\mathcal{V}_h \equiv \{v \in \mathcal{V} \mid v|_K \in \mathbb{P}^2, \ \forall K \in \mathcal{T}_h\}. \tag{3.9}$$

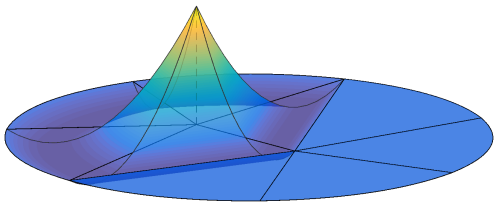
An example of a function that belongs to the space is shown in Figure 3.5(b). We may compare the piecewise linear and quadratic functions shown in Figures 3.2 and 3.5(b), respectively, and



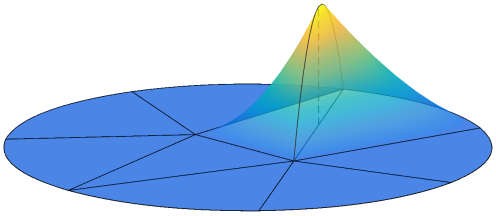
(a) triangulation



(b) example function



(c) vertex shape function ϕ_3



(d) edge shape function ϕ_{20}

Figure 3.5: $\mathbb{P}^2$ finite element space.

intuitively draw a conclusion that the $\mathbb{P}^2$ space provides a better approximation of smooth functions. This intuition is in fact true; we will make a more precise mathematical argument in a later lecture. Like the $\mathbb{P}^1$ space, the $\mathbb{P}^2$ space can be refined by splitting some or all of elements such that $\mathcal{V}_h \subset \mathcal{V}_{h'}$. A successive refinements yield a sequence of approximation spaces $\mathcal{V}_{h_1} \subset \mathcal{V}_{h_2} \subset \cdots \subset \mathcal{V}_{h_n}$ for $h_1 > h_2 > \cdots > h_n$. In fact, a finite element approximations based on piecewise polynomial spaces of degree greater than 1, which include the $\mathbb{P}^2$ space, are often referred to as a *higher-order approximation*, because the solution converges more rapidly with h than for the $\mathbb{P}^1$ space.

Figures 3.5(c) and 3.5(d) show two of the nodal shape functions for the $\mathbb{P}^2$ space associated with the triangulation 3.5(a). Unlike the nodal shape functions for the $\mathbb{P}^1$ space which are associated with only vertices of the triangles, the nodal shape functions for the $\mathbb{P}^2$ space are associated with either vertices or edges.

As another example, suppose the domain of interest Ω is a line in $\mathbb{R}^1$, a square in $\mathbb{R}^2$, a cube in $\mathbb{R}^3$, or any shape that can be mapped to these shapes. Then we can also consider an approximation space consists of global polynomials

$$\mathcal{V}_p \equiv \{v \in \mathcal{V} \mid v \in (\mathbb{P}^p(\Omega))^d\}.$$

In this case, the approximation space can be refined by increasing the polynomial degree; we readily observe that $\mathcal{V}_p \subset \mathcal{V}_{p'}$ for $p \leq p'$. The Galerkin finite element method based on a sequence of global polynomial spaces, $\mathcal{V}_{p_1} \subset \mathcal{V}_{p_2} \subset \cdots \subset \mathcal{V}_{p_n}$ for $p_1 \leq p_2 \leq \cdots \leq p_n$, is called the *spectral method*. The polynomials used in the spectral methods are of very high degree; polynomial spaces of degrees of 100 and higher are routinely used.

3.9 Summary

We summarize key points of this lecture:

1. A triangulation $\mathcal{T}_h$ is a collection of non-overlapping elements $\{K_i\}_{i=1}^{n_e}$ that covers the domain Ω.

2. The act of constructing a triangulation for a given domain is called mesh generation. Mesh generation is a non-trivial task, but there are many open-source and commercial meshers.

3. An approximation space $\mathcal{V}_h$ is a finite-dimensional subspace of $\mathcal{V} \subset H^1(\Omega)$; the space comprises, for example, piecewise polynomials associated with the triangulation.

4. Given a basis $\{\phi_i\}_{i=1}^n$ for $\mathcal{V}_h$, any function $v \in \mathcal{V}_h$ can be identified with a unique coefficients $\hat{v} \in \mathbb{R}^n$, which are the global degrees of freedom of $\mathcal{V}_h$.

5. An approximation space can be successively refined to yield a sequence of approximation spaces.

6. If a nodal basis is used for $H_h^1(\Omega)$, then a subspace $\mathcal{V}_h \subset H_h^1(\Omega)$ that satisfies the essential boundary conditions can be formed by removing nodal shape functions on the closure of the Dirichlet boundary.

7. The Galerkin finite element method solves the variational problem in a finite-dimensional approximation space $\mathcal{V}_h \subset \mathcal{V}$.

8. Given a basis $\{\phi_i\}_{i=1}^n$ of $\mathcal{V}_h$, the coefficients $\hat{u}_h \in \mathbb{R}^n$ associated with the solution $u_h \in \mathcal{V}_h$ solves a $n \times n$ linear system $\hat{A}_h \hat{u}_h = \hat{f}_h$, where $A_{h,ij} = a(\phi_j, \phi_i)$ and $f_{h,i} = \ell(\phi_i)$.

9. If the bilinear form is coercive and continuous and the linear form is continuous, the solution to the Galerkin finite element problem exists and is unique.

10. If the bilinear form is symmetric, coercive, and continuous, then the finite element solution $u_h \in \mathcal{V}_h$ can also be obtained from the minimization principle.

11. The Galerkin finite element procedure can accommodate as its approximation space, for instance, piecewise higher-order polynomials (i.e., higher-order method) or high-order global polynomials (i.e., spectral method).

Lecture 4

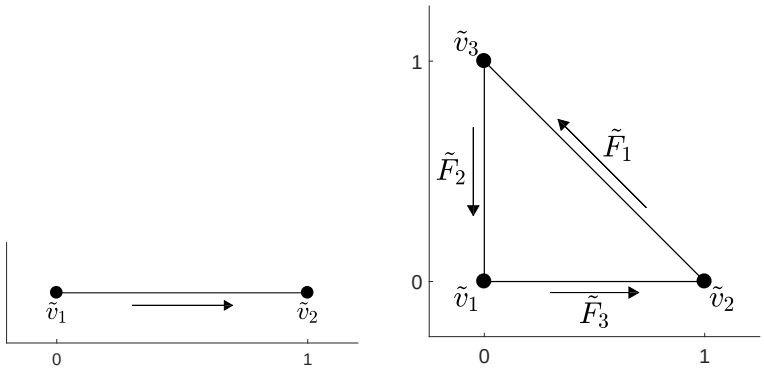
Finite element method: implementation

4.1 Introduction

In this lecture, we introduce technical ingredients required to implement the finite element method. This lecture is organized as follows:

- Section 4.2 introduces a few common finite elements defined on reference domains. Techniques to generate finite elements will be discussed in the section.
- Section 4.3 introduces physical elements defined on a triangulation $\mathcal{T}_h$ and the associated approximation space $\mathcal{V}_h \subset \mathcal{V}$. Techniques to map reference elements to physical elements will be discussed in the section.
- Section 4.4 introduces the concept of numerical quadrature, which we use to evaluate integrals that appear in bilinear and linear forms.
- Section 4.5 discusses the assembly of stiffness matrix and load vector using the ingredients discussed in the preceding sections.
- Section 4.6 discusses the treatment of surface integral terms associated with natural boundary conditions.
- Section 4.7 describes convenient implementation of essential boundary conditions, which are explicitly enforced by the choice of the space.
- Section 4.8 provides a brief discussion of efficient implementation using the BLAS (basic linear algebra subprograms).

This is a rather large lecture that covers significant amount of materials. But, by the end of the lecture, we will have all technical ingredients required to implement a finite element solver.



(a) reference line segment $\tilde{I}$ (b) reference triangle $\tilde{T}$

Figure 4.1: Reference line segment and triangle.

4.2 Reference elements

4.2.1 Reference domains

We first introduce reference domains on which reference finite elements are defined. The first reference domain we introduce is a *reference line segment* $\tilde{I} \subset \mathbb{R}^1$. (Note that all quantities associated with the reference space bear a tilde $(\tilde{\cdot})$.) While the definition of a reference line segment is not universal, our reference line segment, as shown in Figure 4.1(a), is a unit line segment delineated by two vertices

$$\tilde{v}_1 \equiv 0 \quad \text{and} \quad \tilde{v}_2 \equiv 1.$$

(In literature, it is just as common to see a reference line segment defined as $(-1, 1)$.) We consider the line segment oriented in the sense that it points from $\tilde{v}_1$ to $\tilde{v}_2$.

We next introduce a *reference triangle* $\tilde{T} \subset \mathbb{R}^2$. While the definition of a reference triangle is again not universal, our reference triangle, as shown in Figure 4.1(b), is a right triangle delineated by three vertices

$$\tilde{v}_1 \equiv (0, 0), \quad \tilde{v}_2 \equiv (1, 0), \quad \text{and} \quad \tilde{v}_3 \equiv (0, 1).$$

The vertices are ordered counterclockwise, starting with the first vertex at the origin. We also denote the three *facets* of the triangle by

$$\tilde{F}_1 \equiv (\tilde{v}_2, \tilde{v}_3), \quad \tilde{F}_2 \equiv (\tilde{v}_3, \tilde{v}_1), \quad \text{and} \quad \tilde{F}_3 \equiv (\tilde{v}_1, \tilde{v}_2).$$

(A facet is a $d-1$ entity associated with a canonical shape; for a triangle, a facet is an edge.) We choose the convention that the facet number is the same as the vertex number of the vertex on the other side of the triangle. Each facet is oriented such that the collection of the three edges defines the triangle in the counterclockwise orientation.

We could introduce other reference domains including, for instance, a square in $\mathbb{R}^2$, a tetrahedron in $\mathbb{R}^3$, or a cube in $\mathbb{R}^3$; however, in this lecture, we will consider only the reference line segment $\tilde{I}$ and the reference triangle $\tilde{T}$.

Figure 4.2: Linear Lagrange finite element on the reference line segment.

4.2.2 Linear Lagrange finite element on a line segment

We introduce arguably the simplest finite element: linear Lagrange elements on the reference line segment $\tilde{I} \equiv (0,1) \subset \mathbb{R}^1$. To this end, we introduce *Lagrange shape functions* (or *Lagrange basis functions* or *nodal shape functions*) for the space of linear functions on $\tilde{I}$, $\mathbb{P}^1(\tilde{I})$. We choose for our interpolation nodes $\{\tilde{z}_1, \tilde{z}_2\}$ the endpoints of the line segment:

$$\tilde{z}_1 \equiv 0 \quad \text{and} \quad \tilde{z}_2 \equiv 1,$$

as shown in Figure 4.2. Our shape functions are linear functions $\{\tilde{\phi}_1, \tilde{\phi}_2\}$ that satisfy the interpolation condition

$$\tilde{\phi}_i(z_j) = \delta_{ij}, \quad i,j = 1,2; \tag{4.1}$$

here δ_{ij} is the *Kronecker delta* such that $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. We readily confirm that the set of two linear function $\{\tilde{\phi}_1, \tilde{\phi}_2\}$ that satisfies the interpolation condition (4.1) is a basis for $\mathbb{P}^1(\tilde{I})$.

While the linear Lagrange shape functions can be found by inspection, we here follow a more systematic procedure to construct shape functions that generalizes to higher dimensions and higher-order polynomials. To find the basis, we first express the shape functions in terms of the monomial basis $\{1, \tilde{x}\}$:

$$\tilde{\phi}_j(\tilde{x}) = c_1^{(j)} + c_2^{(j)}\tilde{x}, \quad j = 1,2. \tag{4.2}$$

We now apply the interpolation condition (4.1) to find the coefficients. For instance, $\tilde{\phi}_1$ must satisfy

$$\begin{pmatrix} 1 & \tilde{z}_1 \\ 1 & \tilde{z}_2 \end{pmatrix} \begin{pmatrix} c_1^{(1)} \\ c_2^{(1)} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

We can also pose a single matrix equation for the monomial coefficients of both shape functions:

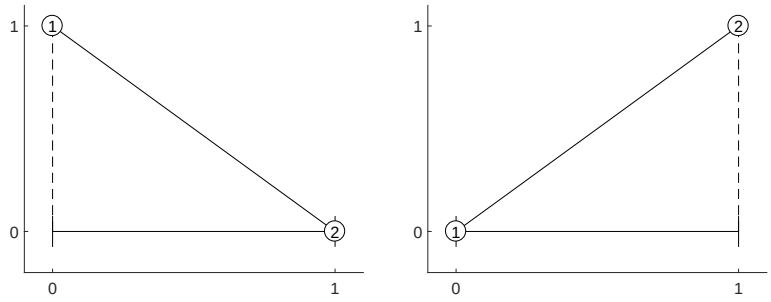
$$\underbrace{\begin{pmatrix} 1 & \tilde{z}_1 \\ 1 & \tilde{z}_2 \end{pmatrix}}_{V} \underbrace{\begin{pmatrix} c_1^{(1)} & c_1^{(2)} \\ c_2^{(1)} & c_2^{(2)} \end{pmatrix}}_{C} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We note that the matrix V is the *Vandermonde matrix* associated with our monomial basis $\{1, \tilde{x}\}$ evaluated at the Lagrange interpolation points $\{\tilde{z}_1, \tilde{z}_2\}$. The matrix V is non-singular as long as the interpolation points are distinct, which is the case for our line segment. The unique coefficients are given by

$$C = V^{-1} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$$

and the associated shape functions are

$$\tilde{\phi}_1(\tilde{x}) = 1 - \tilde{x}_1,$$
$$\tilde{\phi}_2(\tilde{x}) = \tilde{x}_2.$$



(a) $\tilde{\phi}_1$ (b) $\tilde{\phi}_2$

Figure 4.3: Linear Lagrange shape functions on the line segment $\tilde{I}$.

The shape functions are shown in Figure 4.3.

Once we find the coefficients of the shape functions, we can evaluate the value of the functions at any point over $\tilde{I} \subset \mathbb{R}^1$ by evaluating (4.2). We can also differentiate (4.2) to obtain the derivative of the shape functions:

$$\left.\frac{\partial \tilde{\phi}_i}{\partial \tilde{x}}\right|_{\tilde{x}} = c_2^{(i)}, \quad i = 1, 2.$$

More explicitly,

$$\left.\frac{d\tilde{\phi}_1}{d\tilde{x}}\right|_{\tilde{x}} = -1 \quad \text{and} \quad \left.\frac{d\tilde{\phi}_2}{d\tilde{x}}\right|_{\tilde{x}} = 1.$$

The derivatives are constant over the element because the shape functions are linear.

Given the basis $\{\tilde{\phi}_j\}_{j=1}^2$ for $\mathbb{P}^1(\tilde{I})$, we can uniquely associate any $\tilde{v} \in \mathbb{P}^1(\tilde{I})$ with a vector $\hat{\tilde{v}} \in \mathbb{R}^2$:

$$\tilde{v} = \sum_{j=1}^{2} \hat{\tilde{v}}_j \tilde{\phi}_j = \sum_{j=1}^{2} \tilde{v}(\tilde{z}_j)\tilde{\phi}_j$$

We recognize $\hat{\tilde{v}} \in \mathbb{R}^2$ as the *degrees of freedom* with which we can describe functions in $\mathbb{P}^1(\tilde{I})$. For nodal shape functions, $\hat{\tilde{v}}_j = \tilde{v}(\tilde{z}_j)$, $j = 1, 2$, due to the Lagrange interpolation condition; the degrees of freedom are the values of the function at the nodes.

Before we introduce other finite elements, we use the linear Lagrange element as an example to describe three properties that formally defines a *finite element*:

1. the domain over which the element is defined; e.g., the reference line segment $\tilde{I}$.

2. the finite-dimensional linear space of functions; e.g., the linear polynomial space $\mathbb{P}^1(\tilde{I})$.

3. the degrees of freedom used to describe functions; e.g., for $\tilde{v} \in \mathbb{P}^1(\tilde{I})$, the degrees of freedom are the values at the nodes $\{\tilde{v}(\tilde{z}_1), \tilde{v}(\tilde{z}_2)\}$.

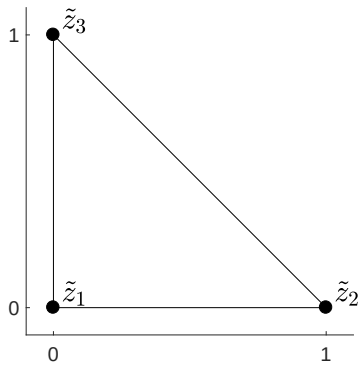


Figure 4.4: Linear Lagrange finite element on the reference triangle.

4.2.3 Linear Lagrange finite element on a triangle

We next introduce a linear Lagrange element on the reference triangle $\tilde{T} \subset \mathbb{R}^2$. Linear functions in $\mathbb{R}^2$ takes the form $a_1 + a_2\tilde{x}_1 + a_3\tilde{x}_2$ and has three degrees of freedom; we hence need to identify a linear independent set of three linear functions. In our case, we wish to identify a set of three linear *Lagrange basis functions* for the space. We choose for our interpolation nodes the three vertices of the triangle

$$\tilde{z}_1 \equiv (0,0), \quad \tilde{z}_2 \equiv (1,0), \quad \text{and} \quad \tilde{z}_3 \equiv (0,1),$$

as shown in Figure 4.4. Our shape functions are linear functions $\{\tilde{\phi}_1, \tilde{\phi}_2, \tilde{\phi}_3\}$ that satisfy the interpolation condition

$$\tilde{\phi}_i(\tilde{z}_j) = \delta_{ij} \quad i,j = 1,\ldots,3, \tag{4.3}$$

where δ_{ij} is the Kronecker delta.

We identify the shape functions using the same procedure used to identify the linear Lagrange shape functions on the unit line segment in Section 4.2.2. We first express the shape functions in terms of the monomial basis $\{1, \tilde{x}_1, \tilde{x}_2\}$:

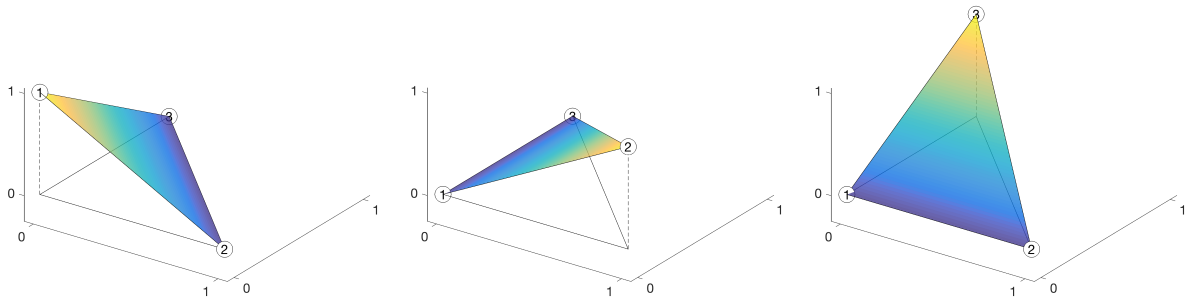
$$\tilde{\phi}_j(\tilde{x}) = c_1^{(j)} + c_2^{(j)}\tilde{x}_1 + c_3^{(j)}\tilde{x}_2 \quad j = 1,2,3. \tag{4.4}$$

We then apply the interpolation condition (4.3) to find the coefficients:

$$\underbrace{\begin{pmatrix} 1 & \tilde{z}_{1,1} & \tilde{z}_{1,2} \\ 1 & \tilde{z}_{2,1} & \tilde{z}_{2,2} \\ 1 & \tilde{z}_{3,1} & \tilde{z}_{3,2} \end{pmatrix}}_{\equiv V} \underbrace{\begin{pmatrix} c_1^{(1)} & c_1^{(2)} & c_1^{(3)} \\ c_2^{(1)} & c_2^{(2)} & c_2^{(3)} \\ c_3^{(1)} & c_3^{(2)} & c_3^{(3)} \end{pmatrix}}_{\equiv C} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $\tilde{z}_{i,j}$ is the j-th coordinate of the i-th interpolation node. The Vandermonde matrix V is non-singular as long as the interpolation points are not collinear, which is equivalent to the condition that the triangle have a finite area; the condition is obviously satisfied for our reference triangle $\tilde{T}$. The coefficients are given by

$$C = V^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix};$$



(a) $\tilde{\phi}_1$ (b) $\tilde{\phi}_2$ (c) $\tilde{\phi}_3$

Figure 4.5: Linear Lagrange shape functions on the reference triangle $\tilde{T}$.

the associated shape functions are

$$\begin{aligned}\tilde{\phi}_1(\tilde{x}) &= 1 - \tilde{x}_1 - \tilde{x}_2 \\ \tilde{\phi}_2(\tilde{x}) &= \tilde{x}_1 \\ \tilde{\phi}_3(\tilde{x}) &= \tilde{x}_2.\end{aligned} \tag{4.5}$$

Figure 4.5 visualizes the three basis functions.

We can also differentiate (4.4) to obtain the gradient of the shape functions:

$$\left.\frac{\partial\tilde{\phi}_j}{\partial\tilde{x}_1}\right|_{\tilde{x}} = c_2^{(j)} \quad \text{and} \quad \left.\frac{\partial\tilde{\phi}_j}{\partial\tilde{x}_2}\right|_{\tilde{x}} = c_3^{(j)}, \quad j = 1, 2, 3.$$

More explicitly,

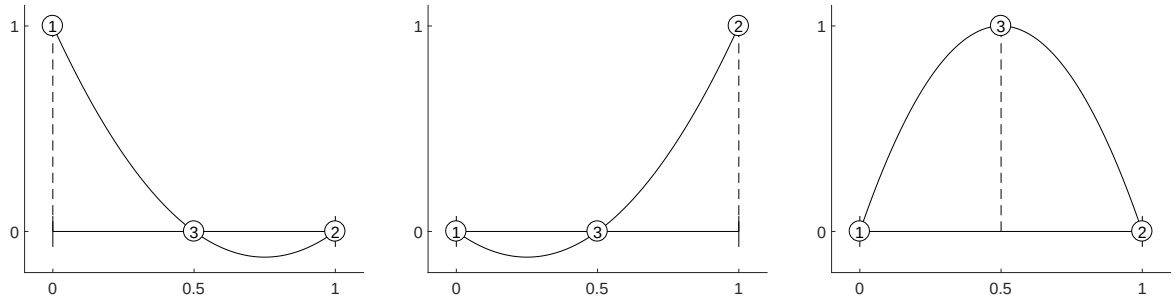
$$\begin{aligned}\left.\frac{\partial\tilde{\phi}_1}{\partial\tilde{x}_1}\right|_{\tilde{x}} &= -1 \quad \text{and} \quad \left.\frac{\partial\tilde{\phi}_1}{\partial\tilde{x}_2}\right|_{\tilde{x}} = -1 \\ \left.\frac{\partial\tilde{\phi}_2}{\partial\tilde{x}_1}\right|_{\tilde{x}} &= 1 \quad \text{and} \quad \left.\frac{\partial\tilde{\phi}_2}{\partial\tilde{x}_2}\right|_{\tilde{x}} = 0 \\ \left.\frac{\partial\tilde{\phi}_3}{\partial\tilde{x}_1}\right|_{\tilde{x}} &= 0 \quad \text{and} \quad \left.\frac{\partial\tilde{\phi}_3}{\partial\tilde{x}_2}\right|_{\tilde{x}} = 1.\end{aligned}$$

For the linear Lagrange element, the derivatives are constant (and trivially spans $\mathbb{P}^0(\tilde{T})$).

Given the basis $\{\tilde{\phi}_j\}_{j=1}^3$, we can uniquely associate any function $\tilde{v} \in \mathbb{P}^1(\tilde{T})$ with a vector $\hat{\tilde{v}} \in \mathbb{R}^3$:

$$\tilde{v} = \sum_{j=1}^{3} \hat{\tilde{v}}_j \tilde{\phi}_j = \sum_{j=1}^{3} \tilde{v}(\tilde{z}_j)\tilde{\phi}_j.$$

Again, for the nodal shape functions, the degrees of freedom are the values of he functions at the nodes, $\{\tilde{v}(\tilde{z}_j)\}_{j=1}^3$. To summarize, our linear Lagrange finite element on a triangle is formally defined by (i) the domain — the reference triangle $\tilde{T}$ —, (ii) the linear function space — the polynomial space $\mathbb{P}^1(\tilde{T})$ —, and (iii) the degrees of freedom — the values at the nodes $\{\tilde{z}_1, \tilde{z}_2, \tilde{z}_3\}$.



(a) $\tilde{\phi}_1$ (b) $\tilde{\phi}_2$ (c) $\tilde{\phi}_3$

Figure 4.6: Quadratic Lagrange shape functions on line segment $\tilde{I}$.

4.2.4 Quadratic Lagrange finite element on a line segment

We now introduce a quadratic Lagrange element on the reference line segment $\tilde{I} \subset \mathbb{R}^1$. A quadratic function in $\mathbb{P}^2(\tilde{I})$ takes the form $a_1 + a_2\tilde{x} + a_3\tilde{x}^2$ and has three degrees of freedom; we hence wish to identify a linearly independent set of three quadratic Lagrange shape functions. To this end, we choose for our Lagrange interpolation nodes the two endpoints and the midpoint,

$$\tilde{z}_1 = 0, \quad \tilde{z}_2 = 1, \quad \text{and} \quad \tilde{z}_3 = 1/2.$$

The ordering of the nodes for a quadratic element is not universal in the finite element literature; we here adhere to the convention that preserves the location of $\tilde{z}_1$ and $\tilde{z}_2$ from the linear element. To find the Lagrange shape functions, we first express the functions in terms of the monomial basis $\{1, \tilde{x}, \tilde{x}^2\}$:

$$\tilde{\phi}_j(\tilde{x}) = c_1^{(j)} + c_2^{(j)}\tilde{x} + c_3^{(j)}\tilde{x}^2, \quad j = 1, 2, 3. \tag{4.6}$$

We then express the interpolation condition $\tilde{\phi}_j(\tilde{z}_i) = \delta_{ij}$ as a 3×3 system $VC = I$, where $C \in \mathbb{R}^{3\times 3}$ is the coefficient matrix so that $C_{ij} = c_i^{(j)}$ and the i-th row of the Vandermonde matrix $V \in \mathbb{R}^{3\times 3}$ is

$$V_{i:} = \begin{pmatrix} 1 & \tilde{z}_i & \tilde{z}_i^2 \end{pmatrix}.$$

The matrix V is non-singular because the monomial basis functions are linearly independent and the three interpolation points are distinct. The shape fucntions are shown in Figure 4.6. The differentiation of 4.6 yields the derivatives of the shape functions:

$$\left.\frac{d\tilde{\phi}_j}{d\tilde{x}}\right|_{\tilde{x}} = c_2^{(j)} + 2c_3^{(j)}\tilde{x}, \quad j = 1, 2, 3.$$

Because the shape functions are quadratic, the derivatives vary linearly over the domain $\tilde{I}$, and they span $\mathbb{P}^1(\tilde{I})$. The degrees of freedom for our nodal $\mathbb{P}^2$ finite element on $\tilde{I}$ are the values of functions at the three nodes $\{\tilde{z}\}_{j=1}^3$.

4.2.5 Quadratic Lagrange finite element on a triangle

We now introduce a quadratic Lagrange finite element on the reference triangle $\tilde{T} \subset \mathbb{R}^2$. A quadratic function in $\mathbb{P}^2(\tilde{T})$ takes the form $a_1 + a_2\tilde{x}_1 + a_3\tilde{x}_2 + a_4\tilde{x}_1^2 + a_5\tilde{x}_1\tilde{x}_2 + a_6\tilde{x}_2^2$ and has six degrees of

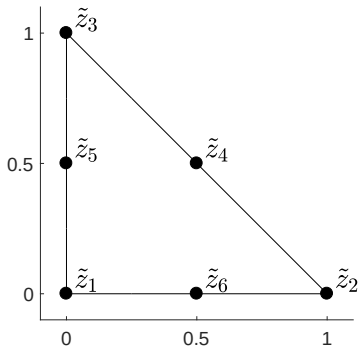


Figure 4.7: Quadratic Lagrange finite element on the reference triangle.

freedom; we hence wish to identify a linearly independently set of six quadratic Lagrange shape functions. To this end, we choose for our Lagrange interpolation nodes the three vertices of the triangle and three points at the middle of the edges,

$$\tilde{z}_1 = (0,0), \quad \tilde{z}_2 = (1,0), \quad \tilde{z}_3 = (0,1), \quad \tilde{z}_4 = (1/2,1/2), \quad \tilde{z}_5 = (0,1/2), \quad \tilde{z}_6 = (1/2,0),$$

as shown in Figure 4.7. The ordering of the nodes is not universal in the finite element literature; we here adhere to the convention that preserves the location of the $\tilde{z}_1$, $\tilde{z}_2$, and $\tilde{z}_3$ from the linear element and, for $i \in \{4,5,6\}$, the $\tilde{z}_i$ is on the midpoint of the $(i-3)$-th edge of the reference triangle. To find the Lagrange shape functions, we first express the basis functions in terms of the monomial basis $\{1, \tilde{x}_1, \tilde{x}_2, \tilde{x}_1^2, \tilde{x}_1\tilde{x}_2, \tilde{x}_2^2\}$:

$$\tilde{\phi}_j(\tilde{x}) = c_1^{(j)} + c_2^{(j)}\tilde{x}_1 + c_3^{(j)}\tilde{x}_2 + c_4^{(j)}\tilde{x}_1^2 + c_5^{(j)}\tilde{x}_1\tilde{x}_2 + c_6^{(j)}\tilde{x}_2^2, \quad j = 1,\ldots,6. \tag{4.7}$$

We then express the interpolation condition $\tilde{\phi}_j(\tilde{z}_i) = \delta_{ij}$ as a 6×6 matrix system $VC = I$, where $C \in \mathbb{R}^{6\times 6}$ is the coefficient matrix so that $C_{ij} = c_i^{(j)}$ and the i-th row of the Vandermonde matrix $V \in \mathbb{R}^{6\times 6}$ is

$$V_{i:} = \begin{pmatrix} 1 & \tilde{z}_{i,1} & \tilde{z}_{i,2} & (z_{i,1})^2 & \tilde{z}_{i,1}\tilde{z}_{i,2} & (\tilde{z}_{i,2})^2 \end{pmatrix}.$$

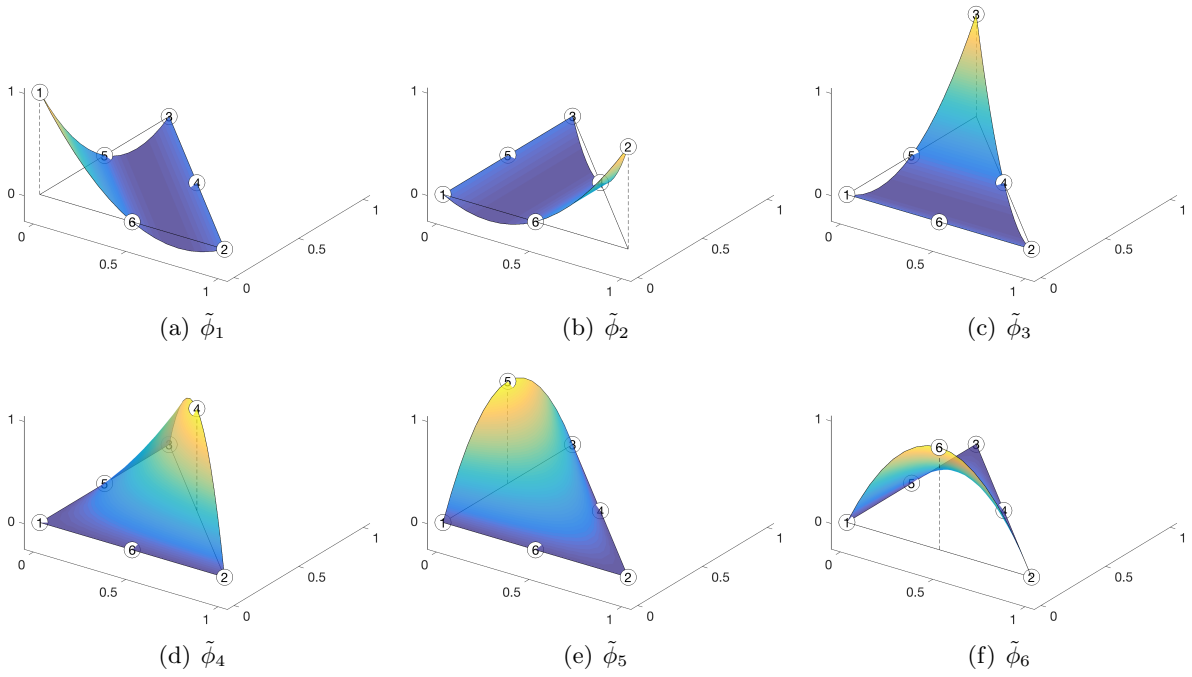
The matrix V is non-singular, and the linear system has a unique solution: $C = V^{-1}$. Figure 4.8 shows the six basis functions. The differentiation of (4.7) yields the gradient of the shape functions,

$$\left.\frac{\partial\tilde{\phi}_j}{\partial\tilde{x}_1}\right|_{\tilde{x}} = c_2^{(j)} + 2c_4^{(j)}\tilde{x}_1 + c_5^{(j)}\tilde{x}_2$$
$$\left.\frac{\partial\tilde{\phi}_j}{\partial\tilde{x}_2}\right|_{\tilde{x}} = c_3^{(j)} + c_5^{(j)}\tilde{x}_1 + 2c_6^{(j)}\tilde{x}_2.$$

For the quadratic Lagrange element, the derivatives are linear functions, and they span $\mathbb{P}^1(\tilde{T})$.

4.2.6 Generalization: an advanced method using Legendre polynomials

We can generalize the procedure discussed so far in this section to generate Lagrange shape functions of an arbitrary degree on an arbitrary domain $\tilde{K}$. Say we wish to generate Lagrange shape functions



(a) $\tilde{\phi}_1$ (b) $\tilde{\phi}_2$ (c) $\tilde{\phi}_3$

(d) $\tilde{\phi}_4$ (e) $\tilde{\phi}_5$ (f) $\tilde{\phi}_6$

Figure 4.8: Quadratic Lagrange shape functions on the reference triangle.

for a polynomial space of degree p with a dimension n_s and the interpolation nodes $\{\tilde{z}_i\}_{i=1}^{n_s}$. We first identify *any* basis $\{\tilde{\psi}_k\}_{k=1}^{n_s}$. We then express the Lagrange shape functions as

$$\tilde{\phi}_j(\tilde{x}) = \sum_{k=1}^{n_s} c_k^{(j)} \psi_k(\tilde{x}), \quad \tilde{x} \in \tilde{K}, \quad j = 1, \ldots, n_s. \tag{4.8}$$

The coefficients that satisfy the interpolation condition $\tilde{\phi}_j(\tilde{z}_i) = \delta_{ij}$ must satisfy the $n_s \times n_s$ system

$$\begin{pmatrix} \psi_1(\tilde{z}_1) & \cdots & \psi_{n_s}(\tilde{z}_1) \\ \vdots & \ddots & \vdots \\ \psi_1(\tilde{z}_{n_s}) & \cdots & \psi_{n_s}(\tilde{z}_{n_s}) \end{pmatrix} \begin{pmatrix} c_1^{(1)} & \cdots & c_1^{(n_s)} \\ \vdots & \ddots & \vdots \\ c_{n_s}^{(1)} & \cdots & c_{n_s}^{(n_s)} \end{pmatrix} = I_{n_s}, \tag{4.9}$$

where I_{n_s} is the $n_s \times n_s$ identity matrix. The derivative of the shape functions are then given by

$$\left.\frac{\partial \tilde{\phi}_j}{\partial \tilde{x}_i}\right|_{\tilde{x}} = \sum_{k=1}^{n_s} c_k^{(j)} \left.\frac{\partial \tilde{\psi}_k}{\partial \tilde{x}_i}\right|_{\tilde{x}}, \quad \tilde{x} \in \tilde{K}. \tag{4.10}$$

Note that this is a generalization, or simply an abstraction, of the method we have discussed for linear and quadratic shape functions on a line segment and triangle discussed in the previous sections.

The efficient implementation of this procedure relies on the efficient evaluation of the values and gradients of the basis functions $\{\tilde{\psi}_k\}_{k=1}^{n_s}$. One convenient choice for $\{\tilde{\psi}_k\}_{k=1}^{n_s}$ are the *Legendre polynomials*, which is a set of orthogonal polynomials in $L^2(\tilde{I})$. We provide a formal definition.

Definition 4.1 (Legendre polynomial)**.** The Legendre polynomials $\{\psi_i\}_{i=0}^n$ are hierarchical polynomials defined on a unit line segment $\tilde{I} \equiv (0,1)$ such that

(i) $\text{span}\{\psi_i\}_{i=0}^n = \mathbb{P}^n(\tilde{I})$

(ii) $(\psi_i, \psi_j)_{L^2(\Omega)} \equiv \int_{\tilde{I}} \psi_i \psi_j dx = \delta_{ij}$, $\forall i, j = 1, \ldots, n$.

The orthogonality condition (ii) implies that the set $\{\psi_i\}_{i=0}^n$ is linearly independent and hence is a basis for $\mathbb{P}^n(\tilde{I})$. In addition the Legendre polynomials of arbitrary degree and their respective derivatives can be evaluated using recurrence relations. We can hence use (4.8) and (4.10) with the Legendre polynomials as the underlying basis. In higher dimensions, we may use the tensor product of Legendre polynomials as the underlying basis. The use of the Legendre polynomial as the underlying polynomials also ensures the linear system (4.9) remains well-posed even if the polynomial degree is very high; the Vandermonde matrix associated with the monomials become ill-conditioned for a high-degree polynomials as the monomials becomes nearly linearly dependent.

4.3 Physical elements

4.3.1 Geometry mapping

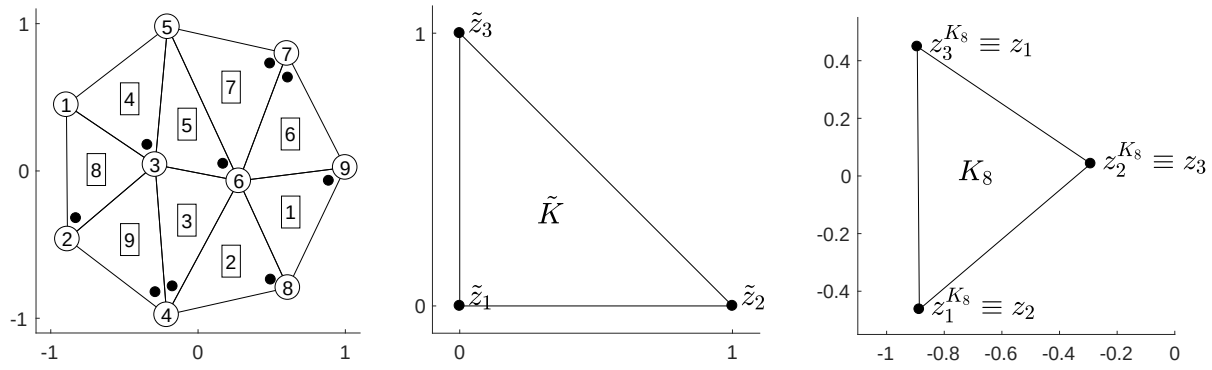
We have so far introduced shape functions $\{\tilde{\phi}_i\}_{i=1}^{n_s}$ defined on a reference element $\tilde{K}$, where the reference element may be the reference line segment $\tilde{I}$ or the reference triangle $\tilde{T}$. We now wish to construct a set of shape functions $\{\phi_i\}_{i=1}^n$ which spans the approximation space $\mathcal{V}_h \subset \mathcal{V}$. To clearly distinguish between quantities defined on the reference domain and those defined on the actual physical domain, we will qualify the latter quantities with the adjective *physical*. (Note that the quantities associated with the physical space, unlike those associated with the reference space, do not bear a tilde $(\tilde{\cdot})$.)

To begin, we create a mapping from a point $\tilde{x}$ in the reference element $\tilde{K}$ to a point x in the physical element K. The physical element is delineated by n_s nodes, $\{z_\alpha^K\}_{\alpha=1}^{n_s}$. To map a point $\tilde{x} \in \tilde{K}$ in the reference domain to a point $x \in K$ in the physical domain, we employ a geometry mapping, $\mathcal{G}^K : \tilde{K} \to K$ given by

$$x = \mathcal{G}^K(\tilde{x}) \equiv \sum_{\alpha=1}^{n_s} z_\alpha^K \tilde{\phi}_\alpha(\tilde{x}), \tag{4.11}$$

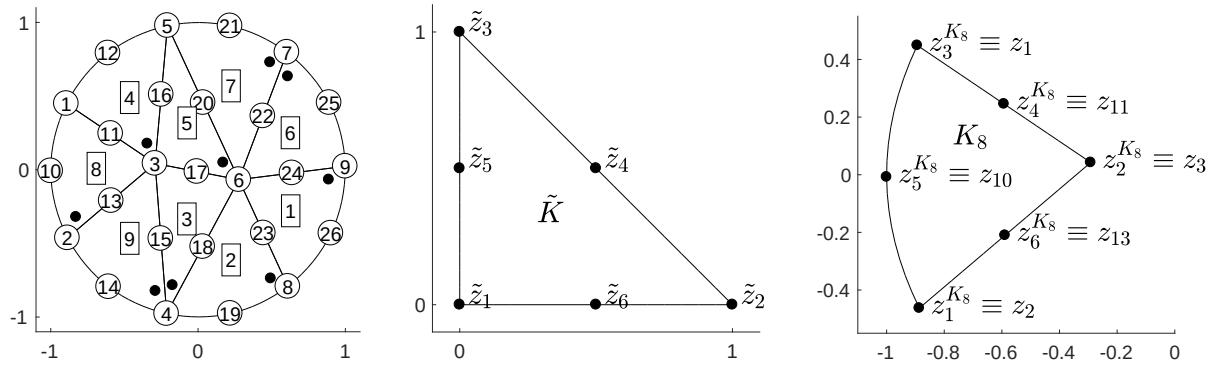
where $z_\alpha^K \in \mathbb{R}^d$ is the coordinates of the α-th node of the physical element K, $\tilde{\phi}_\alpha \in \mathbb{P}^p(\tilde{K})$ is the Lagrange shape function associated with the α-th node of the reference element, and n_s is the number of the shape functions. Our geometry mapping $\mathcal{G}^K : \tilde{K} \to K$ is a unique map that (i) is a polynomial map of degree p and (ii) maps the Lagrange interpolation points $\{\tilde{z}_\alpha\}_\alpha^{n_s}$ of the reference element $\tilde{K}$ to the respective Lagrange interpolation points $\{z_\alpha\}_\alpha^{n_s}$ of the physical element K. The geometry mapping $\mathcal{G}^K : \tilde{K} \to K$ is invertible under a reasonable condition; we will discuss the condition shortly.

An example of a $\mathbb{P}^1$ geometry mapping of a triangle is shown in Figure 4.9. The physical domain K_8 is defined by the physical nodes $\{z_1^{K_8}, z_2^{K_8}, z_3^{K_8}\}$ which are in turn identified by the three nodes $\{z_2, z_3, z_1\}$ of the mesh. Because the reference element is linear, the mapping $\mathcal{G}^{K_8} : \tilde{K} \to K_8$ is affine.



(a) triangulation (b) reference domain $\tilde{K} \equiv \tilde{T}$ (c) physical domain K_8

Figure 4.9: $\mathbb{P}^1$ geometry mapping.



(a) triangulation (b) reference domain $\tilde{K} \equiv \tilde{T}$ (c) physical domain K_8

Figure 4.10: $\mathbb{P}^2$ geometry mapping.

As an another example, we provide a $\mathbb{P}^2$ geometry mapping of a triangle in Figure 4.10. The physical domain K_8 is defined by the physical nodes $\{z_1^{K_8}, z_2^{K_8}, z_3^{K_8}, z_4^{K_8}, z_5^{K_8}, z_6^{K_8}\}$ which are in turn identified by the six nodes $\{z_2, z_3, z_1, z_{11}, z_{10}, z_{13}\}$ of the mesh. Because the reference element is quadratic, the mapping $\mathcal{G}^{K_8} : \tilde{K} \to K_8$ is also quadratic. We can represent curved geometries using $\mathbb{P}^{p>1}$ geometry mapping; as we will see later, the accurate representation of the curved geometry is important to realize higher-order approximations of boundary value problems.

We now introduce a few quantities derived from the mapping. We can differentiate the mapping (4.11) to evaluate the *Jacobian* $J^K : \tilde{K} \to \mathbb{R}^{d \times d}$ given by

$$J_{ij}^K(\tilde{x}) \equiv \left.\frac{\partial x_i}{\partial \tilde{x}_j}\right|_{\tilde{x}} = \left.\frac{\partial \mathcal{G}_i^K}{\partial \tilde{x}_j}\right|_{\tilde{x}} = \sum_{\alpha=1}^{n_s} z_{\alpha,i}^K \left.\frac{\partial \tilde{\phi}_\alpha}{\partial \tilde{x}_j}\right|_{\tilde{x}},$$

where $z_{\alpha,i}^K$ is the i-th coordinate of the α-th node of element K. The Jacobian characterizes how an infinitesimal line segment $d\tilde{l}$ at $\tilde{x} \in \tilde{K}$ is mapped to an infinitesimal line segment dl at $x \in K$; specifically, $dl = J^K(\tilde{x}) d\tilde{l}$. If the mapping $\mathcal{G}^K : \tilde{K} \to K$ is a polynomial of degree p, the Jacobian

$J^K : \tilde{K} \to \mathbb{R}^{d\times d}$ is a polynomial of degree $p-1$. For a $\mathbb{P}^1$ mapping, the Jacobian is constant over $\tilde{K}$.

The determinant of the Jacobian $\det(J^K(\tilde{x}))$, or more compactly $|J^K(\tilde{x})|$, relates the reference area $d\tilde{x}$ to the physical area dx by

$$dx = |J^K(\tilde{x})| d\tilde{x}.$$

If the mapping $\mathcal{G}^K : \tilde{K} \to K$ is a polynomial of degree p, the Jacobian $J^K : \tilde{K} \to \mathbb{R}^{d\times d}$ is a polynomial of degree $p-1$, and the determinant of the Jacobian $|J^K| : \tilde{K} \to \mathbb{R}$ is a polynomial of degree $d(p-1)$. For a $\mathbb{P}^1$ mapping, the determinant of the Jacobian is constant over $\tilde{K}$.

Now we consider the *inverse mapping* $(\mathcal{G}^K)^{-1} : K \to \tilde{K}$, which maps a physical point $x \in K$ to a reference point $\tilde{x} \in \tilde{K}$. The inverse mapping exists for all points if and only if

$$|J^K(\tilde{x})| > 0 \quad \forall \tilde{x} \in \tilde{K}. \tag{4.12}$$

For a general polynomial mapping $\mathcal{G}^K : \tilde{K} \to K$ of degree p, this condition must be checked for all $\tilde{x} \in \tilde{K}$. Moreover, the function $(\mathcal{G}^K)^{-1} : K \to \tilde{K}$ is in general not a polynomial because the inverse of a polynomial is not a polynomial. Consequently, the evaluation of the inverse mapping requires the solution of a nonlinear problem: given $x \in K$, find $\tilde{x} \in \tilde{K}$ such that $\mathcal{G}^K(\tilde{x}) = x$.

The inverse mapping is greatly simplified for a $\mathbb{P}^1$ mapping $\mathcal{G}^K : \tilde{K} \to K$ because the Jacobian $J^K : \tilde{K} \to K$ is constant. First, for a $\mathbb{P}^1$ mapping, the condition (4.12) is equivalent to the condition that (i) the area of the physical triangle is finite and (ii) the vertices are ordered in the counterclockwise manner. (Note that if the vertices are not ordered in the counterclockwise manner, then the mapping is inverted and we would obtain a negative area.) For a $\mathbb{P}^1$ mapping, $|J^K|$ is constant and $|J^K|/2$ is the area of the physical triangle; the factor of $1/2$ is needed because the area of our reference triangle $\tilde{T}$ is $1/2$. Second, because $\mathcal{G}^K : \tilde{K} \to K$ is linear, $(\mathcal{G}^K)^{-1} : K \to \tilde{K}$ is also linear; given $x \in K$, we can solve a linear system to find $\tilde{x} \in \tilde{K}$.

We can also compute the Jacobian associated with the inverse mapping, or the *inverse Jacobian*, $(J^K)^{-1} : \tilde{K} \to \mathbb{R}^{d\times d}$:

$$\left.\frac{\partial \tilde{x}_i}{\partial x_j}\right|_{\tilde{x}} = \left.\frac{\partial((\mathcal{G}^K)^{-1})_i}{\partial x_j}\right|_{\tilde{x}} = ((J^K(\tilde{x}))^{-1})_{ij}.$$

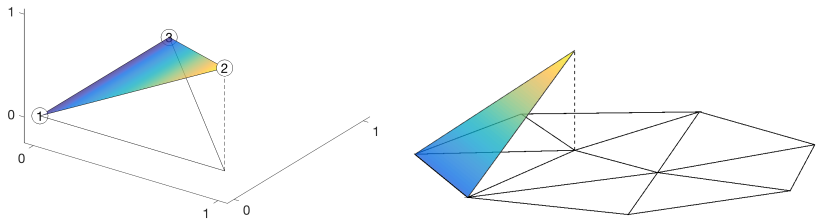
The algebraic inverse of the Jacobian $J^K = \frac{\partial x}{\partial \tilde{x}}$ is the inverse Jacobian $(J^K)^{-1} = \frac{\partial \tilde{x}}{\partial x}$. The inverse Jacobian $\frac{\partial x_i}{\partial \tilde{x}_j}$ is well defined at $\tilde{x} \in \tilde{K}$ if and only if $|J^K(\tilde{x})| > 0$. For a general polynomial mapping $\mathcal{G}^K : \tilde{K} \to K$ of degree p, the inverse Jacobian is not a polynomial because the inverse of a polynomial is in general not a polynomial. For a $\mathbb{P}^1$ mapping, the inverse Jacobian $(J^K)^{-1} : K \to \mathbb{R}^{d\times d}$ is constant because the Jacobian $J^K : \tilde{K} \to \mathbb{R}^{d\times d}$ is constant.

4.3.2 Physical shape functions

Given a geometry mapping $\mathcal{G}^K : \tilde{K} \to K$, we now introduce physical shape functions $\{\phi_\alpha^K\}_{\alpha=1}^{n_s}$ associated with the physical element $K \in \mathcal{T}_h$. We choose the basis functions that satisfy

$$\phi_\alpha^K(x = \mathcal{G}^K(\tilde{x})) = \tilde{\phi}_\alpha(\tilde{x}) \quad \forall \tilde{x} \in \tilde{K}, \quad \alpha = 1, \ldots, n_s, \tag{4.13}$$

where $\tilde{x} \mapsto x = \mathcal{G}^K(\tilde{x})$ is provided by the geometry mapping (4.11). In words, the physical basis function ϕ_α^K evaluated at the physical point $x(\tilde{x}) \in K$ takes the same value as the associated reference basis function ϕ_α evaluated at the associated reference point $\tilde{x} \in \tilde{K}$.



(a) reference shape function $\tilde{\phi}_2$ (b) physical shape function $\phi_2^{K_9}$

Figure 4.11: Mapping of shape functions for a $\mathbb{P}^1$ element.

We can also differentiate (4.13) to obtain the derivative of the physical basis functions in the physical space: given any $\tilde{x} \in \tilde{K}$,

$$\left.\frac{\partial \phi_\alpha^K}{\partial x_i}\right|_{x=\mathcal{G}^K(\tilde{x})} = \sum_{j=1}^{d} \left.\frac{\partial \tilde{x}_j}{\partial x_i}\right|_{\tilde{x}} \left.\frac{\partial \tilde{\phi}_\alpha}{\partial \tilde{x}_j}\right|_{\tilde{x}}, \quad i = 1, \ldots, d, \ \alpha = 1, \ldots, n_s. \tag{4.14}$$

The relationship may be expressed more explicitly using matrices and vectors; for $d = 2$,

$$\begin{pmatrix} \frac{\partial \phi_\alpha^K}{\partial x_1} \\ \frac{\partial \phi_\alpha^K}{\partial x_2} \end{pmatrix}_{x=\mathcal{G}^K(\tilde{x})} = \begin{pmatrix} \frac{\partial \tilde{x}_1}{\partial x_1} & \frac{\partial \tilde{x}_2}{\partial x_1} \\ \frac{\partial \tilde{x}_1}{\partial x_2} & \frac{\partial \tilde{x}_2}{\partial x_2} \end{pmatrix}_{\tilde{x}} \begin{pmatrix} \frac{\partial \tilde{\phi}_\alpha}{\partial \tilde{x}_1} \\ \frac{\partial \tilde{\phi}_\alpha}{\partial \tilde{x}_2} \end{pmatrix}_{\tilde{x}}.$$

Or, noting that $\frac{\partial \tilde{x}_j}{\partial x_i} = ((J^K)^{-1})_{ji} = ((J^K)^{-T})_{ij}$, a more compact expression (for any d) is

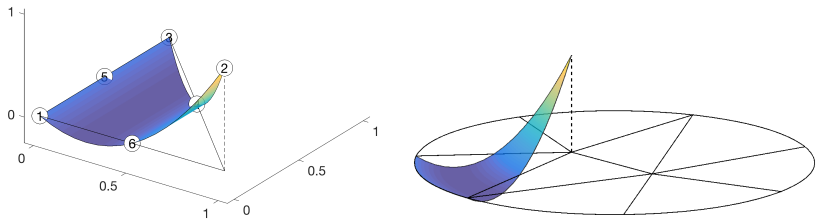
$$\nabla \phi_\alpha^K(x = \mathcal{G}^K(\tilde{x})) = (J^K)^{-T}(\tilde{x}) \tilde{\nabla} \tilde{\phi}_\alpha(\tilde{x}). \tag{4.15}$$

The expressions (4.13) and (4.14) (or equivalently (4.15)) allow us to evaluate the value and gradient, respectively, of the physical shape function ϕ_α^K at a physical point $x = \mathcal{G}^K(\tilde{x}) \in K$ associated with a select reference point $\tilde{x} \in \tilde{K}$; we will soon see this is exactly the capability we need to evaluate stiffness matrices and load vectors.

As an example, consider the physical element K_9 in the triangulation that comprises linear elements shown in Figure 4.9(a). The element K_9 is delineated by the nodes $\{z_1^{K_9} = z_4, z_2^{K_9} = z_3, z_3^{K_9} = z_2\}$. The reference shape function $\tilde{\phi}_2 \in \mathbb{P}^1(\tilde{T})$ maps to the physical shape function $\phi_2^{K_9}$ as shown in Figure 4.11. We in fact recognize that $\phi_2^{K_9}$ is the restriction of the physical (global) shape function ϕ_3 associated with z_3; formally, $\phi_2^{K_9} \equiv \phi_3|_{K_9}$.

As another example, consider the physical element K_9 in the triangulation that comprises quadratic elements shown in Figure 4.10(a). The element K_9 is delineated by the nodes $\{z_1^{K_9} = z_4, z_2^{K_9} = z_3, z_3^{K_9} = z_2, z_4^{K_9} = z_{13}, z_5^{K_9} = z_{14}, z_6^{K_9} = z_{15}\}$. The reference shape function $\tilde{\phi}_2 \in \mathbb{P}^2(\tilde{T})$ maps to the physical shape function $\phi_2^{K_9}$ as shown in Figure 4.12. We again recognize that $\phi_2^{K_9}$ is the restriction of the physical (global) shape function ϕ_3 associated with z_3; formally, $\phi_2^{K_9} \equiv \phi_3|_{K_9}$.

We make one remark about our physical basis functions defined by (4.13). Even though the reference basis function $\tilde{\phi} : \tilde{K} \to \mathbb{R}$ is a polynomial in $\tilde{K}$, the physical basis function $\phi_\alpha^K : K \to \mathbb{R}$ is in general not a polynomial in K. To see this, we observe that $\phi_\alpha^K(x) = \tilde{\phi}_\alpha((\mathcal{G}^K)^{-1}(x))$, $\forall x \in K$; because the inverse map $(\mathcal{G}^K)^{-1} : K \to \tilde{K}$ is not a polynomial in K for a $\mathbb{P}^{p>1}$ geometry mapping, the function $\phi_\alpha^K(\cdot) = \tilde{\phi}_\alpha((\mathcal{G}^K)^{-1}(\cdot)) = \tilde{\phi}_\alpha \circ (\mathcal{G}^K)^{-1}$ is not a polynomial in K. Conversely, we note that $\phi_\alpha^K(\mathcal{G}^K(\cdot)) = \phi_\alpha^K \circ \mathcal{G}^K = \tilde{\phi}(\cdot)$ *is* a polynomial in $\tilde{K}$. In short, $\phi_\alpha^K \in \mathbb{P}^p \circ (\mathcal{G}^K)^{-1}$ or equivalently $\phi_\alpha^K \circ \mathcal{G}^K \in \mathbb{P}^p(\tilde{K})$.



(a) reference shape function $\tilde{\phi}_2$ (b) physical shape function $\phi_2^{K_9}$

Figure 4.12: Mapping of shape functions for a $\mathbb{P}^2$ element.

We note each physical element is indeed a finite element characterized by the three properties: the domain is $K \in \mathcal{T}_h$; the finite-dimensional approximation space is $\mathbb{P}^p \circ (\mathcal{G}^K)^{-1}$; the degrees of freedom are the function values at the physical nodes $\{v(z_\alpha^K)\}_{\alpha=1}^{n_s}$. The collection of the finite elements define our approximation space, which can be compactly stated as

$$\mathcal{V}_h \equiv \{v \in H^1(\Omega) \mid v|_K \circ \mathcal{G}^K \in \mathbb{P}^p(\tilde{K}),\ \forall K \in \mathcal{T}_h\}.$$

For a $\mathbb{P}^1$ geometry mapping, the notation is simplified; because the inverse mapping $(\mathcal{G}^K)^{-1} : K \to \tilde{K}$ is affine, the physical shape function $\phi_\alpha^K(\cdot) = \tilde{\phi}_\alpha((\mathcal{G}^K)^{-1}(\cdot))$ is a polynomial in K. Thus, for $\mathbb{P}^1$ geometry mapping,

$$\mathcal{V}_h \equiv \{v \in H^1(\Omega) \mid v|_K \in \mathbb{P}^p(K),\ \forall K \in \mathcal{T}_h\},$$

as introduced in the previous lecture.

Before we conclude this section, we clarify the nomenclature. In this section, we considered physical elements that result from using the same polynomial space for the geometry mapping $\mathcal{G} : \tilde{K} \to K$ and the function representation; these elements are called *isoparametric elements*. In general, the polynomials used for the geometry mapping and function representation need not be the same. An element that uses a higher degree representation of the geometry than functions is called a *superparametric element*. Conversely, an element that use a lower degree representation of the geometry than functions is called a *subparametric element*.

4.4 Numerical quadrature

4.4.1 Motivation

The evaluation of bilinear and linear forms that appear in the weak form of boundary value problems requires integration of functions. This integration is performed by *numerical quadrature* (or just *quadrature*). Specifically, our quadrature problem is as follows: given a reference element $\tilde{K}$ and a function $f : \tilde{K} \to \mathbb{R}$, estimate the integral

$$I \equiv \int_{\tilde{K}} f(\tilde{x}) d\tilde{x}$$

by

$$Q \equiv \sum_{q=1}^{n_q} \tilde{\rho}_q f(\tilde{\xi}_q),$$

where $\{\tilde{\xi}_q \in \tilde{K}\}_{q=1}^{n_q}$ is a set of *quadrature points* and $\{\tilde{\rho}_q \in \mathbb{R}\}_{q=1}^{n_q}$ is the associated set of *quadrature weights*.

4.4.2 Gauss quadrature in $\mathbb{R}^1$

We first consider a one-dimensional quadrature for a unit line segment $\tilde{I} \equiv (0,1) \subset \mathbb{R}^1$. (Note: one-dimensional quadrature rules are more often defined for the line segment $(-1,1)$; in this lecture we define them for $(0,1)$ to be consistent with our definition of a unit line segment.) While there are many different families of one-dimensional quadrature rules, we focus on arguably the most efficient quadrature rule: the Gauss quadrature.

The n_q-point Gauss quadrature rule is defined by quadrature points $\{\tilde{\xi}_q \in \tilde{I}\}_{q=1}^{n_q}$ and quadrature weights $\{\tilde{\rho}_q \in \tilde{I}\}_{q=1}^{n_q}$ such that the rule integrates exactly polynomials of degrees up to and including $2n_q - 1$: i.e.,

$$\int_{\tilde{I}} f(\tilde{x}) d\tilde{x} = \sum_{q=1}^{n_q} \tilde{\rho}_q f(\tilde{\xi}_q) \quad \forall f \in \mathbb{P}^{2n_q-1}(\tilde{I}). \tag{4.16}$$

Our intuition might suggest the existence of such a quadrature rule, as the polynomials of degree $2n_q - 1$ have $2n_q$ degrees of freedom and the n_q-point quadrature rule also has $2n_q$ degrees of freedom — n_q point locations and n_q weigh values.

We can show the existence of a n_q-point quadrature rule that exactly integrates polynomials of degree $2n_q - 1$ in a constructive manner. To this end, we use the scaled Legendre polynomials $\{\tilde{\psi}_i\}$ over $\tilde{I}$ defined in Definition 4.1. We first choose the quadrature points $\{\tilde{\xi}_q\}_{q=1}^{n_q}$ as the roots of the degree n_q Legendre polynomial:

$$\tilde{\psi}_{n_q}(\tilde{\xi}_q) = 0, \quad q = 1, \ldots, n_q. \tag{4.17}$$

We then choose the quadrature weights $\{\tilde{\rho}_q\}_{q=1}^{n_q}$ to satisfy the following linear equation:

$$\begin{pmatrix} \tilde{\psi}_0(\tilde{\xi}_1) & \cdots & \tilde{\psi}_0(\tilde{\xi}_{n_q}) \\ \vdots & \ddots & \vdots \\ \tilde{\psi}_{n_q-1}(\tilde{\xi}_1) & \cdots & \tilde{\psi}_{n_q-1}(\tilde{\xi}_{n_q}) \end{pmatrix} \begin{pmatrix} \tilde{\rho}_1 \\ \vdots \\ \tilde{\rho}_{n_q} \end{pmatrix} = \begin{pmatrix} \int_{\tilde{I}} \tilde{\psi}_0(\tilde{x}) d\tilde{x} \\ \vdots \\ \int_{\tilde{I}} \tilde{\psi}_{n_q-1}(\tilde{x}) d\tilde{x} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.18}$$

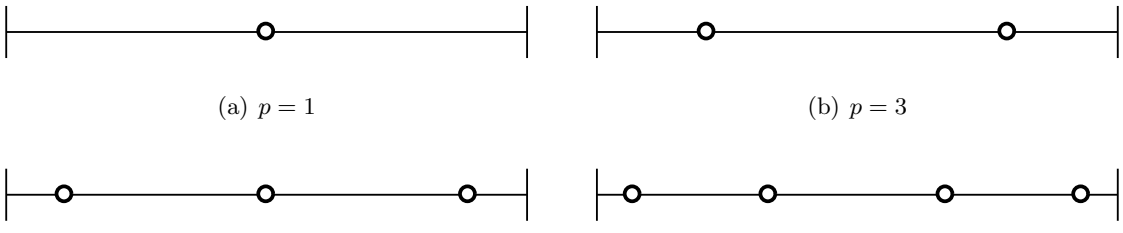
We wish to show the conditions (4.17) and (4.18) yield a quadrature rule that integrates exactly polynomials of degree $2n_q - 1$. To begin, we introduces a basis $\{p_i\}_{i=0}^{2n_q-1}$ for $\mathbb{P}^{2n_q-1}(\tilde{I})$ such that $\forall \tilde{x} \in \tilde{I}$

$$\begin{aligned} p_0(\tilde{x}) &= \tilde{\psi}_0(\tilde{x}) = 1, & p_1(\tilde{x}) &= \tilde{\psi}_1(\tilde{x}), & \ldots, \quad p_{n_q-1}(\tilde{x}) &= \tilde{\psi}_{n_q-1}(\tilde{x}) \\ p_{n_q}(\tilde{x}) &= \tilde{\psi}_{n_q}(\tilde{x})\tilde{\psi}_0(\tilde{x}), & p_{n_q+1}(\tilde{x}) &= \tilde{\psi}_{n_q}(\tilde{x})\tilde{\psi}_1(\tilde{x}), & \ldots, \quad p_{2n_q-1}(\tilde{x}) &= \tilde{\psi}_{n_q}(\tilde{x})\tilde{\psi}_{n_q-1}(\tilde{x}). \end{aligned}$$

The polynomials $\{p_i\}_{i=0}^{2n_q-1}$ is a basis for $\mathbb{P}^{2n_q-1}(\tilde{I})$; the set spans $\mathbb{P}^{2n_q-1}(\tilde{I})$ and is linearly independent because p_i is a polynomial of degree (exactly) i for $i = 0, \ldots, 2n_q - 1$. We now wish to confirm that

$$\int_{\tilde{I}} p_i(\tilde{x}) d\tilde{x} = \sum_{q=1}^{n_q} \tilde{\rho}_q p_i(\tilde{\xi}_q), \quad \forall i = 0, \ldots, 2n_q - 1, \tag{4.19}$$

which is equivalent to the original condition (4.16). We first readily confirm that the first n_q basis functions, $\{p_i\}_{i=0}^{n_q} \equiv \{\tilde{\psi}_i\}_{i=0}^{n_q}$, are integrated exactly because our weights $\{\tilde{\rho}_q\}_{q=1}^{n_q}$ are chosen to



(a) $p = 1$ (b) $p = 3$

(c) $p = 5$ (d) $p = 7$

Figure 4.13: Gauss quadrature points for $(0, 1)$.

p	n_q	$\tilde{\xi}$	$\tilde{\rho}$
1	1	0.500000000000000	1.000000000000000
3	2	0.211324865405187	0.500000000000000
		0.788675134594813	0.500000000000000
5	3	0.112701665379258	0.277777777777778
		0.500000000000000	0.444444444444444
		0.887298334620742	0.277777777777778
7	4	0.069431844202974	0.173927422568727
		0.330009478207572	0.326072577431273
		0.669990521792428	0.326072577431273
		0.930568155797026	0.173927422568727

Table 4.1: Gauss quadrature rules on $\tilde{I}$ for $p = 1$ to 7 polynomials.

integrate the functions in condition (4.18). To prove that the next n_q basis functions, $\{p_i\}_{i=n_q}^{2n_q-1}$ are integrated exactly, we observe that the left-hand side of 4.19 for $i = n_q, \ldots, 2_{n_q} - 1$ yields

$$(\text{LHS}) = \int_{\tilde{I}} p_{n_q+j}(\tilde{x}) d\tilde{x} = \int_{\tilde{I}} \tilde{\psi}_{n_q}(\tilde{x}) \tilde{\psi}_j(\tilde{x}) d\tilde{x} = 0, \quad j = 0, \ldots, n_q - 1,$$

since the Legendre polynomials are orthogonal in $L^2(\tilde{I})$. On the other hand, the right-hand side of 4.19 for $i = n_q, \ldots, 2_{n_q} - 1$ yields

$$(\text{RHS}) = \sum_{q=1}^{n_q} \tilde{\rho}_q p_{n_q+j}(\tilde{\xi}_q) = \sum_{q=1}^{n_q} \tilde{\rho}_q \tilde{\psi}_{n_q}(\tilde{\xi}_q) \tilde{\psi}_j(\tilde{\xi}_q) = 0, \quad j = 0, \ldots, n_q - 1,$$

since $\{\tilde{\xi}_q\}_{q=1}^{n_q}$ are the roots of $\tilde{\psi}_{n_q}$ by (4.17). In summary, the exact integration condition (4.19) is satisfied (i) for $i = 0, \ldots, n_q - 1$ because of the choice of the weights $\{\tilde{\rho}_q\}_{q=1}^{n_q}$ by (4.18) and (ii) for $i = n_q, \ldots, 2n_q - 1$ because of the choice of the points $\{\tilde{\xi}_q\}_{q=1}^{n_q}$ by (4.17).

Table 4.1 shows the Gauss quadrature rules for $n_q = 1, \ldots, 4$. As visualized in Figure 4.13, the quadrature points are clustered towards the endpoints.

4.4.3 Numerical quadrature in $\mathbb{R}^d$

Similarly to the Gauss quadrature in $\mathbb{R}^1$, there also exist efficient quadrature rules for integration of domains in $\mathbb{R}^{d>1}$. If the domain is a unit square $(0, 1)^2 \equiv \tilde{I}^2 \subset \mathbb{R}^2$ or a unit cube $(0, 1)^3 \equiv \tilde{I}^3 \subset \mathbb{R}^3$,

p	n_q	$\tilde{\xi}_1$	$\tilde{\xi}_2$	$\tilde{\rho}$
1	1	0.333333333333333	0.333333333333333	0.500000000000000
2	3	0.166666666666667	0.166666666666667	0.166666666666667
		0.666666666666667	0.166666666666667	0.166666666666667
		0.166666666666667	0.666666666666667	0.166666666666667
4	6	0.091576213509771	0.091576213509771	0.054975871827661
		0.816847572980459	0.091576213509771	0.054975871827661
		0.091576213509771	0.816847572980459	0.054975871827661
		0.445948490915965	0.445948490915965	0.111690794839006
		0.108103018168070	0.445948490915965	0.111690794839006
		0.445948490915965	0.108103018168070	0.111690794839006

Table 4.2: Efficient quadrature rules on $\tilde{T}$ for $p = 1$ to 4 polynomials.

then we can obtain the associated quadrature rule by the tensor-product of one-dimensional Gauss rules; for example, for $(0,1)^2$,

$$\int_{\tilde{x}_2=0}^{1}\int_{\tilde{x}_1=0}^{1} f(\tilde{x}_1,\tilde{x}_2)d\tilde{x}_1 d\tilde{x}_2 \approx \sum_{i_2=1}^{n_q^{\text{1d}}}\sum_{i_1=1}^{n_q^{\text{1d}}} \tilde{\rho}_{i_2}^{\text{1d}}\tilde{\rho}_{i_1}^{\text{1d}} f(\tilde{\xi}_1^{\text{1d}},\tilde{\xi}_2^{\text{1d}}).$$

These rules maximize the degree of tensor-product polynomials exactly integrated for a given number of quadrature points.

For a domain in $\mathbb{R}^{d<1}$ that does not result from a tensor-product of a one-dimensional domain, the "optimal" quadrature rules are much more difficult to identify. In fact, the optimal rules for (say) a triangle is not as universally standardized as that for a square. Table 4.2 shows examples of efficient quadrature rule for our unit right triangle, which exactly integrates polynomials of degree p. Figure 4.14 visualizes the quadrature points. Similarly to the one-dimensional Gauss rule, the quadrature points are clustered towards the edge of the triangular domain.

4.5 Assembly

4.5.1 Local stiffness matrices and vectors

We now put together the techniques we have learned in this lecture to evaluate local stiffness matrices and load vectors. To provide a concrete example, in this section we consider $\mathcal{V} \equiv H^1(\Omega)$, a bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ such that

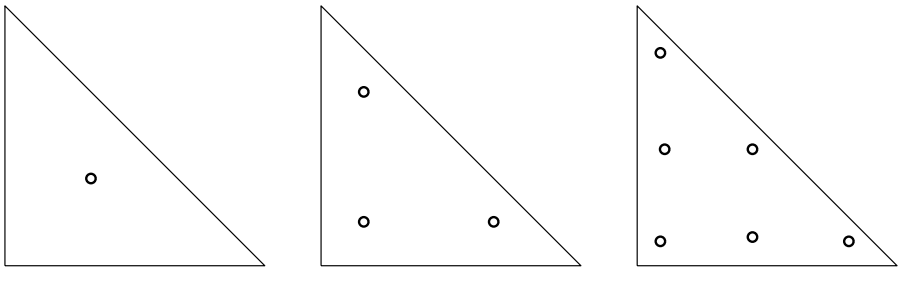
$$a(w,v) \equiv \int_\Omega \nabla v \cdot a \nabla w dx, \quad \forall w, v \in \mathcal{V}, \tag{4.20}$$

for $a : \Omega \to \mathbb{R}^{d\times d}$ the diffusion tensor field, and a linear form $\ell : \mathcal{V} \to \mathbb{R}$ such that

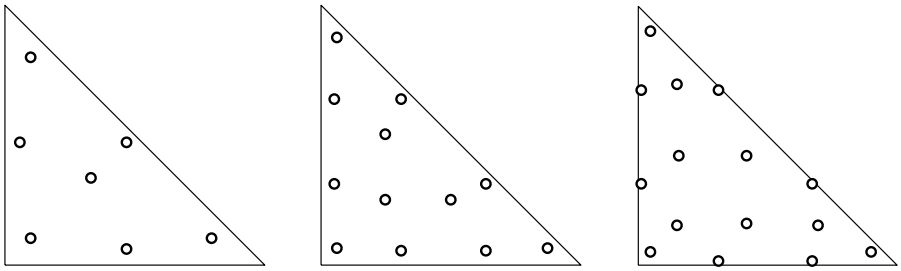
$$\ell(v) \equiv \int_\Omega v f dx \quad \forall v \in \mathcal{V}, \tag{4.21}$$

for $f : \Omega \to \mathbb{R}$ the source function. Our approximation space is given by

$$\mathcal{V}_h \equiv \{v \in \mathcal{V} \mid v|_K \circ \mathcal{G}^K \in \mathbb{P}^p(\tilde{K}),\ \forall K \in \mathcal{T}_h\}, \tag{4.22}$$



(a) $p = 1$ (b) $p = 2$ (c) $p = 4$



(d) $p = 5$ (e) $p = 6$ (f) $p = 7$

Figure 4.14: Numerical quadrature points for the reference triangle.

where $\mathcal{G}^K : \tilde{K} \to K$ is a $\mathbb{P}^p$ geometry mapping. The number of shape functions per element is denoted by n_s.

We first consider the evaluation of the *local load vector* $\hat{f}^K \in \mathbb{R}^{n_s}$ such that

$$\hat{f}^K_\alpha \equiv \ell(\phi^K_\alpha), \quad \alpha = 1, \ldots, n_s.$$

For our model linear form (4.21), the local load vector is given by

$$\hat{f}^K_\alpha \equiv \ell(\phi^K_\alpha) \equiv \int_K \phi^K_\alpha(x) f(x) dx, \quad \alpha = 1, \ldots, n_s.$$

We now change the integration domain from the physical element K to the reference element $\tilde{K}$; this is in preparation for the application of the numerical quadrature on the reference element. We employ (i) the geometry mapping for the argument of f, $x = \mathcal{G}^K(\tilde{x})$, (ii) the relationship between the reference and physical shape functions, $\phi^K_\alpha(x \equiv \mathcal{G}^K(\tilde{x})) = \tilde{\phi}_\alpha(\tilde{x})$, and (iii) the area transformation $dx = |J^K(\tilde{x})| d\tilde{x}$, where $|J^K|$ is the determinant of the Jacobian of the geometry mapping, to obtain

$$\hat{f}^K_\alpha = \int_{\tilde{K}} \phi_\alpha(\tilde{x}) f(\mathcal{G}^K(\tilde{x})) |J(\tilde{x})| d\tilde{x}. \tag{4.23}$$

We finally apply a n_q-point numerical quadrature to "evaluate" the integral:

$$\hat{f}^K_\alpha \text{ "=" } \sum_{q=1}^{n_q} \tilde{\rho}_q \tilde{\phi}_\alpha(\tilde{\xi}_q) f(\mathcal{G}^K(\tilde{\xi}_q)) |J(\tilde{\xi}_q)|.$$

Here, we put the equality in quotes because the evaluation is exact only if the integrand is a polynomial that can be integrated exactly using the quadrature rule. Otherwise, the use of the numerical quadrature results in an *approximation* rather than an *evaluation*. In any event, we can readily evaluate the quantities in the summand using the techniques discussed in the previous sections. We here make a listing for a reference: (i) quadrature rule $\{\tilde{\xi}_q, \tilde{\rho}_q\}_{q=1}^{n_q}$ was discussed in Sections 4.4; (ii) the evaluation of quantities associated with geometry mapping, $\mathcal{G}^K$ and J^K, was discussed in Sections 4.3.1; (iii) the evaluation of reference shape functions $\{\tilde{\phi}_\alpha\}_{\alpha=1}^{n_s}$ was discussed in Section 4.2.

We now consider the evaluation of the *local stiffness matrix* $\hat{A}^K \in \mathbb{R}^{n_s \times n_s}$ such that

$$\hat{A}^K_{\alpha,\beta} \equiv a(\phi^K_\beta, \phi^K_\alpha) \quad \forall \alpha, \beta = 1, \ldots, n_s.$$

For our model bilinear form (4.20), the local stiffness matrix is given by

$$\hat{A}^K_{\alpha,\beta} \equiv a(\phi^K_\beta, \phi^K_\alpha) = \int_K \nabla \phi^K_\alpha(x) \cdot a(x) \nabla \phi^K_\beta(x) dx, \quad \alpha, \beta = 1, \ldots, n_s.$$

Following the same procedure we used for the evaluation (or approximation) of the local load vector $\hat{f}^K_h \in \mathbb{R}^{n_s}$, we now change the integration domain from the physical element K to the reference element $\tilde{K}$. We employ (i) the geometry mapping for the argument of a, $x = \mathcal{G}^K(\tilde{x})$, (ii) the relationship between the derivatives of the reference and physical shape functions, $\nabla \phi^K(x = \mathcal{G}(\tilde{x})) = J^K(\tilde{x})^{-T} \tilde{\nabla} \tilde{\phi}(\tilde{x})$, and (iii) the area transformation $dx = |J^K(\tilde{x})| d\tilde{x}$, where $|J^K|$ is the determinant of the Jacobian of the geometry mapping, to obtain

$$\hat{A}^K_{\alpha,\beta} = \int_{\tilde{K}} (J^K(\tilde{x})^{-T} \tilde{\nabla} \tilde{\phi}_\alpha(\tilde{x})) \cdot a(\mathcal{G}^K(\tilde{x})) J^K(\tilde{x})^{-T} \tilde{\nabla} \tilde{\phi}_\beta(\tilde{x}) |J(\tilde{x})| d\tilde{x} \tag{4.24}$$

We finally apply a n_q-point numerical quadrature to "evaluate" the integral:

$$\hat{A}^K_{\alpha,\beta} \text{“=”} \sum_{q=1}^{n_q} \tilde{\rho}_q (J^K(\tilde{\xi}_q)^{-T} \tilde{\nabla} \tilde{\phi}_\alpha(\tilde{\xi}_q)) \cdot a(\mathcal{G}^K(\tilde{\xi}_q)) J^K(\tilde{\xi}_q)^{-T} \tilde{\nabla} \tilde{\phi}_\beta(\tilde{\xi}_q) |J^K(\tilde{\xi}_q)|.$$

Again, we put the equality in quotes because the evaluation is exact only if the integrand is a polynomial that can be integrated exactly using the quadrature rule. In particular, in the presence of a curved element, the integrand is almost always non-polynomial because the inverse Jacobian is non-polynomial. The techniques used to evaluate the element stiffness matrix are almost identical to those used to evaluate the element load vector, except this time we also employ the technique to evaluate the derivative of the physical shape functions discussed in Section 4.3.2. If the bilinear form contains other terms, such as a convection term of the form $\int_\Omega v b \cdot \nabla w dx$ or a reaction term of the form $\int_\Omega c v w dx$, they can also be evaluated in a similar manner.

We make one note about the expressions for the local load vector and stiffness matrix. In literature, the expressions (4.23) and (4.24) are sometimes restated as

$$\hat{f}^K_\alpha = \int_{\tilde{K}} \tilde{\phi}_\alpha(\tilde{x}) \underbrace{f(\mathcal{G}^K(\tilde{x})) |J^K(\tilde{x})|}_{\equiv f^{\text{trans}}(\tilde{x})} d\tilde{x},$$

$$\hat{A}^K_{\alpha,\beta} = \int_{\tilde{K}} \tilde{\nabla} \tilde{\phi}_\alpha(\tilde{x}) \cdot \underbrace{J^K(\tilde{x})^{-1} a(\mathcal{G}^K(\tilde{x})) J^K(\tilde{x})^{-T} |J^K(\tilde{x})|}_{\equiv a^{\text{trans}}(\tilde{x})} \tilde{\nabla} \tilde{\phi}_\beta(\tilde{x}) d\tilde{x},$$

where $f^{\text{trans}} : \tilde{K} \to \mathbb{R}$ and $a^{\text{trans}} : \tilde{K} \to \mathbb{R}^{d\times d}$ are the transformed coefficients for the reference element $\tilde{K}$. Note that the difference lies only in the interpretation; the local load vectors and stiffness matrices for both formulations are mathematically identical.

4.5.2 Global matrix and vector assembly

We have so far introduced technical ingredients required to assemble, for any $K \in \mathcal{T}_h$, the local load vector $\hat{f}^K \in \mathbb{R}^{n_s}$ such that

$$\hat{f}^K_\alpha = \ell(\phi^K_\alpha), \quad \alpha = 1, \dots, n_s,$$

and a local stiffness matrix $\hat{A}^K \in \mathbb{R}^{n_s\times n_s}$ such that

$$\hat{A}^K_{\alpha,\beta} = a(\phi^K_\beta, \phi^K_\alpha), \quad \alpha, \beta = 1, \dots, n_s.$$

We now wish to assemble the local stiffness matrices and vectors construct the (global) stiffness matrix and vector.

To this end, we employ the element-to-node connectivity map

$$\theta_{K\text{-}n} : \{1, \dots, n_e\} \times \{1, \dots, n_s\} \to \{1, \dots, n\}$$

such that $i = \theta_{K\text{-}n}(k, \alpha)$ is the global node number of the α-th node of the k-th element. We recall that the map is typically stored as a table of the size $n_e \times n_s$. Then, to form the (global) stiffness matrix $\hat{A}_h \in \mathbb{R}^{n\times n}$, we successively insert the local stiffness matrices $\hat{A}^{K_k} \in \mathbb{R}^{n_s\times n_s}$ for $k = 1, \dots, n_e$ according to

$$\hat{A}_{h,ij} \leftarrow \hat{A}_{h,ij} + \hat{A}^{K_k}_{\alpha\beta}, \quad \alpha, \beta = 1, \dots, n_s,$$

for $i = \theta_{K\text{-}n}(k, \alpha)$ and $j = \theta_{K\text{-}n}(k, \beta)$. Similarly, to form the (global) load vector $\hat{f}_h \in \mathbb{R}^n$, we successively insert the local load vectors $\hat{f}^{K_k} \in \mathbb{R}^{n_s}$ for $k = 1, \dots, n_e$ according to

$$\hat{f}_{h,i} \leftarrow \hat{f}_{h,i} + \hat{f}^{K_k}_\alpha, \quad \alpha = 1, \dots, n_s,$$

for $i = \theta_{K\text{-}n}(k, \alpha)$.

4.6 Natural boundary conditions

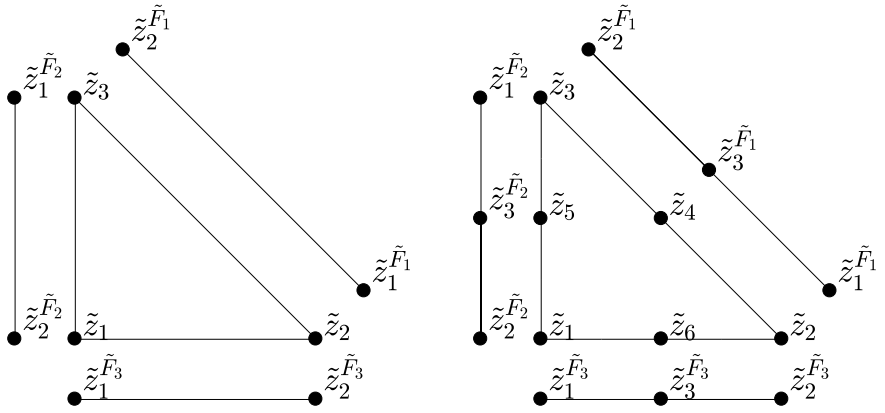
4.6.1 Reference facet-to-node maps

In this section, we introduce technical tools required to evaluate surface integral terms in natural boundary conditions; i.e., Neumann and Robin boundary conditions. To this end, we first consider a linear triangular reference element and introduce a mapping that relates nodes on a facet to nodes on the element. Figure 4.15(a) shows the relationship between the nodes. Our facet-to-node map is

$$\theta_{\tilde{F}\text{-}n} : \{1, 2, 3\} \times \{1, 2\} \to \{1, 2, 3\}$$

such that

$$\begin{aligned} &\theta_{\tilde{F}\text{-}n}(1,1) = 2, && \theta_{\tilde{F}\text{-}n}(1,2) = 3,\\ &\theta_{\tilde{F}\text{-}n}(2,1) = 3, && \theta_{\tilde{F}\text{-}n}(2,2) = 1,\\ &\theta_{\tilde{F}\text{-}n}(3,1) = 1, && \theta_{\tilde{F}\text{-}n}(3,2) = 2. \end{aligned}$$



(a) linear element (b) quadratic element

Figure 4.15: Element-facet relationship for linear and quadratic Lagrange triangular elements.

Note that $j = \theta_{\tilde{F}\text{-}n}(i,k)$ is the Lagrange node on $\tilde{T}$ identified with the k-th Lagrange node of the facet i; i.e.,

$$\tilde{z}_{j \equiv \theta_{\tilde{F}\text{-}n}(i,k)} = \tilde{z}_k^{\tilde{F}_i}, \quad i = 1,2,3,\ k = 1,2.$$

Similarly, we consider a quadratic triangular reference element and introduce a mapping that relates nodes on a facet to nodes on the element. Figure 4.15(b) shows the relationship between the nodes. Our facet-to-node map is

$$\theta_{\tilde{F}\text{-}n} : \{1,2,3\} \times \{1,2,3\} \to \{1,\ldots,6\}$$

such that

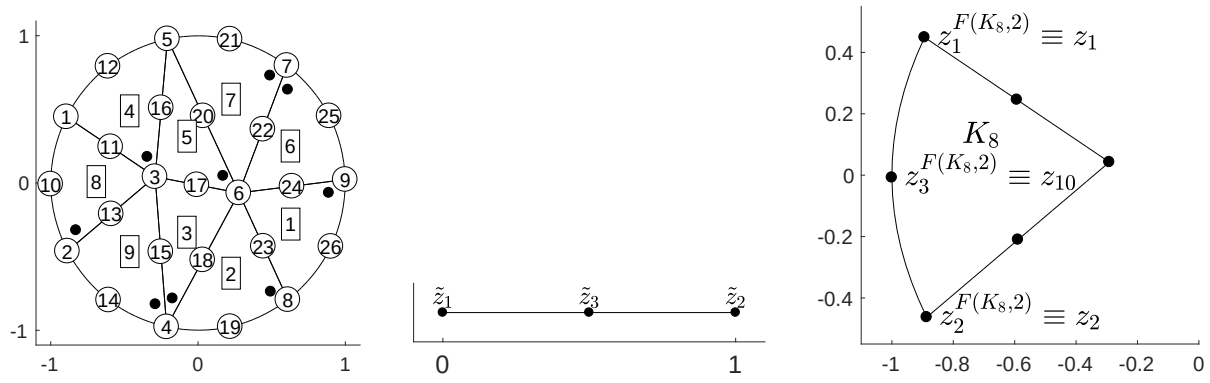
$$\begin{aligned}
&\theta_{\tilde{F}\text{-}n}(1,1) = 2, && \theta_{\tilde{F}\text{-}n}(1,2) = 3, && \theta_{\tilde{F}\text{-}n}(1,3) = 4\\
&\theta_{\tilde{F}\text{-}n}(2,1) = 3, && \theta_{\tilde{F}\text{-}n}(2,2) = 1, && \theta_{\tilde{F}\text{-}n}(2,3) = 5\\
&\theta_{\tilde{F}\text{-}n}(3,1) = 1, && \theta_{\tilde{F}\text{-}n}(3,2) = 2, && \theta_{\tilde{F}\text{-}n}(3,3) = 6.
\end{aligned}$$

As before, $j = \theta_{\tilde{F}\text{-}n}(i,k)$ is the Lagrange node on $\tilde{T}$ identified with the k-th Lagrange node of the facet i; i.e.,

$$\tilde{z}_{j \equiv \theta_{\tilde{F}\text{-}n}(i,k)} = \tilde{z}_k^{\tilde{F}_i}, \quad i = 1,2,3,\ k = 1,2,3.$$

4.6.2 Geometry mapping for facets

In Section 4.3.1, we introduced a geometry mapping $\mathcal{G}^K : \tilde{K} \to K$ from a reference element $\tilde{K} \subset \mathbb{R}^d$ to a physical element $K \subset \mathbb{R}^d$. While this is the only mapping we need to evaluate bilinear and linear forms that only require integration over Ω, forms that require integration over boundaries — for example a Neumann boundary $\Gamma_N \subset \partial\Omega$ — requires another mapping. Specifically, we require a mapping from a $(d-1)$-dimensional reference element to the associated physical facet that lies in a d-dimensional space; the physical facet is a $d-1$-dimensional manifold embedded in a d-dimensional space. For concreteness, in this section we restrict ourselves to the case $d = 2$, and



(a) triangulation (b) reference line segment $\tilde{K} \equiv \tilde{I}$ (c) physical facet $F(K_8, 2)$

Figure 4.16: $\mathbb{P}^2$ geometry mapping for a facet.

consider the geometry mapping $\mathcal{G}^F : \tilde{I} \to F$ from the reference line segment $\tilde{I} \subset \mathbb{R}^1$ to a physical facet $F \subset \mathbb{R}^2$ of a triangle.

Figure 4.16 shows a concrete example of an geometry mapping from a reference line segment $\tilde{I}$ to a physical facet F, the second facet of the physical element K_8. We recall from Section 4.6.1 that the Lagrange nodes on the reference facet $\tilde{F}_{i\equiv 2}$ are related to those on the reference element $\tilde{T}$ by $\tilde{z}_\alpha = \tilde{z}_\gamma^{\tilde{F}_{i\equiv 2}}$ for $\alpha = \theta_{\tilde{F}\text{-}n}(i \equiv 2, \gamma)$, $\gamma = 1, 2, 3$; for the case in Figure 4.16, we obtain $\{\tilde{z}_1^{\tilde{F}_2} \equiv \tilde{z}_3, \tilde{z}_2^{\tilde{F}_2} \equiv \tilde{z}_1, \tilde{z}_3^{\tilde{F}_2} \equiv \tilde{z}_5\}$. Accordingly, the Lagrange nodes on the physical facet F are related to those on the physical element $\{z_j^{K_8}\}_{j=1}^6$, and the (global) nodes $\{z_i\}_{i=1}^n$, by $\{z_1^{F)} \equiv z_3^{K_8} \equiv z_1, z_2^F \equiv z_1^{K_8} \equiv z_2, z_3^F \equiv z_5^{K_8} \equiv z_{10}\}$. Our geometry mapping for the facet, $\mathcal{G}^F : \tilde{I} \to F$, is given by

$$x = \mathcal{G}^F(\tilde{s}) \equiv \sum_{\alpha=1}^{n_s^F} z_\alpha^F \tilde{\chi}_\alpha(\tilde{s}), \tag{4.25}$$

where $z_\alpha^F \in \mathbb{R}^{d\equiv 2}$ is the coordinates of the α-th node of the physical facet $F \subset \mathbb{R}^{d\equiv 2}$, $\chi_\alpha \in \mathbb{P}^{p\equiv 2}(\tilde{I})$ is the Lagrange shape function associated with the α-th node of the reference line segment $\tilde{I} \subset \mathbb{R}^1$, and $n_s^F = 3$ for the $\mathbb{P}^2(\tilde{I})$ space. Note that the input is $\tilde{s} \in \tilde{I} \subset \mathbb{R}^{d-1\equiv 1}$ while the output is $x \in F \subset \mathbb{R}^{d\equiv 2}$. Similar to the geometry mapping (4.11) for the element, the geometry mapping (4.25) is a polynomial map of degree p that maps the interpolation nodes $\{\tilde{z}_i\}$ of the reference line segment $\tilde{I}$ to the associated nodes $\{z_i^F\}$ of the physical facet F.

We can differentiate (4.25) to obtain the Jacobian, $J^F : \tilde{I} \to \mathbb{R}^{d\times(d-1)}$, given by

$$J_{ij}^F(\tilde{s}) \equiv \left.\frac{\partial x_i}{\partial \tilde{s}_j}\right|_{\tilde{s}} = \left.\frac{\partial \mathcal{G}_i^F}{\partial \tilde{s}_j}\right|_{\tilde{s}} = \sum_{\alpha=1}^{n_s^F} z_{\alpha,i}^F \left.\frac{\partial \tilde{\phi}_\alpha}{\partial \tilde{s}_j}\right|_{\tilde{s}}, \quad i = 1, \ldots, d,\ j = 1, \ldots, d-1.$$

The Jacobian is rectangular because the mapping is from $\mathbb{R}^{d-1}$ to $\mathbb{R}^d$. The Jacobian $J^F(\tilde{s})$ characterizes how an infinitesimal line segment $d\tilde{l}$ at $\tilde{s} \in \tilde{I}$ maps to an infinitesimal line segment dl at $s \in F$: $dl = J^F(\tilde{s})d\tilde{l}$.

The physical length ds is related to the reference length $d\tilde{s}$ by the relationship

$$ds = |J^F(\tilde{s})|d\tilde{s},$$

where, in two dimensions,

$$|J^F(\tilde{s})| \equiv \sqrt{J_{11}^F(\tilde{s})^2 + J_{21}^F(\tilde{s})^2}.$$

Analogously to the determinant condition $|J(\tilde{x})| > 0 \ \forall \tilde{x} \in \tilde{K}$ for the element mapping, a facet mapping is valid if and only if

$$|J^F(\tilde{s})| > 0 \quad \forall \tilde{s} \in \tilde{I}.$$

This condition is automatically satisfied if $|J^K(\tilde{x})| > 0 \ \forall \tilde{x} \in \bar{\tilde{K}}$ for the element K associated with the facet F.

The evaluation of forms associated with a boundary value problem sometimes also require the evaluation of the *unit outward-pointing normal vector*. In two dimension, the normal vector is given by

$$n^F(\tilde{s}) = \frac{1}{|J^F(\tilde{s})|} \begin{pmatrix} J_{21}^F(\tilde{s}) \\ -J_{11}^F(\tilde{s}) \end{pmatrix}.$$

This definition of the normal vector is for our counterclockwise convention for the facet orientation; the sign needs to be reversed if the clockwise convection is used for the facet orientation.

4.6.3 Local stiffness matrix and vectors of facet terms

We now consider evaluation of a load vector that requires integration on a boundary and, in turn, on a facet. Again to provide a concrete example, in this section we consider $\mathcal{V} \equiv H^1(\Omega)$, a linear form $\ell : \mathcal{V} \to \mathbb{R}$ such that

$$\ell(v) \equiv \int_{\Gamma_N} vg dx \quad \forall v \in \mathcal{V}, \tag{4.26}$$

for $g : \Gamma_N \to \mathbb{R}$ the Neumann source function. Our approximation space $\mathcal{V}_h$ is as defined in (4.22). The number of shape fucntions per element is denoted by n_s, and the number of shape functions on a facet of the element is denoted by n_s^F. (We only consider the load vector, and not the stiffness matrix, as the evaluation procedures are essentially the same.)

By way of preliminaries, we first provide a triangular for the boundary Γ_N. The boundary triangulation

$$\mathcal{T}_h^{\Gamma_N} \equiv \{F_i\}_{i=1}^{n_f}$$

is a set of n_f non-overlapping facets that cover the boundary Γ_N: (i) $F_i \cap F_j = \emptyset$, $i \neq j$, and (ii) $\cup_{i=1}^{n_f} \bar{F}_i = \bar{\Gamma}_N$. We assume that the facet triangulation $\mathcal{T}_h^{\Gamma_N}$ is compatible with the domain triangulation $\mathcal{T}_h \equiv \{K_i\}_{i=1}^{n_e}$ in the sense that each $F_i \in \mathcal{T}_h^{\Gamma_N}$ is a facet of a unique element. Given this compatibility condition, we introduce two mappings:

1. $\theta_{F\text{-}K} : \{1, \ldots, n_f\} \to \{1, \ldots, n_e\}$. Mapping from a physical facet number to the element to which the facet belongs.

2. $\theta_{F\text{-}\tilde{F}} : \{1, \ldots, n_f\} \to \{1, 2, 3\}$. Mapping from a physical facet number to the local facet index.

These two mappings combined implies that a physical facet F is the $i = \theta_{F\text{-}\tilde{F}}(F)$-th facet of the element $K_{j \equiv \theta_{F\text{-}K}(F)}$.

To evaluate the local load vector $\tilde{f}^K \in \mathbb{R}^{n_s}$, we appeal to the fact that the restriction of an element shape function ϕ^K in $\mathbb{R}^d$ to a facet F is a shape function in $\mathbb{R}^{d-1}$:

$$\phi_\alpha^K|_F = \chi_\gamma^F, \quad \gamma = 1, \ldots, n_s^F,$$

for $K = \theta_{F\text{-}K}(F)$, $\alpha = \theta_{\tilde{F}\text{-}n}(i,\gamma)$, and $i = \theta_{F\text{-}\tilde{F}}(F)$. We recall that the mapping $\theta_{\tilde{F}\text{-}n}$ is the mapping from facet nodes to element nodes introduced in Sections 4.2.3 and 4.2.5; $\alpha = \theta_{\tilde{F}\text{-}(n)}(i,\gamma)$ is the element node that corresponds to the γ-th facet node of the i-th facet. Our approach is hence to evaluate the boundary integral using the facet shape functions $\{\chi_\gamma^F\}_{\gamma=1}^{n_s^F}$ and then to map the appropriate integrals to the element integral. Specifically, we note that if the linear form $\ell(\cdot)$ only involves integration on a boundary,

$$\ell(\phi_\alpha^K|_F) = \ell(\chi_\gamma^F), \quad \gamma = 1,\ldots,n_s^F, \tag{4.27}$$

for $K = \theta_{F\text{-}K}(F)$, $\alpha = \theta_{\tilde{F}\text{-}n}(i,\gamma)$, and $i = \theta_{F\text{-}\tilde{F}}(F)$. For our model linear form (4.26) the evaluation of the integral for $\{\chi_\gamma^F\}_{\gamma=1}^{n_s^F}$ yields

$$\hat{g}_\gamma^F \equiv \ell(\chi_\gamma^F) = \int_F \chi_\gamma^F(s) g(s) ds, \quad \gamma = 1,\ldots,n_s^F.$$

To evaluate (or approximate) the integral, we change the integration domain from the physical facet F to the reference line segment $\tilde{I}$. We employ (i) the facet geometric mapping for the argument of g, $s \equiv \mathcal{G}^F(\tilde{s})$, (ii) the relationship between the reference and physical shape functions, $\tilde{\chi}_\gamma^F(s \equiv \mathcal{G}^F(\tilde{s})) = \tilde{\chi}(\tilde{s})$, and (iii) the length transformation $ds = |J^F(\tilde{s})| d\tilde{s}$, to obtain

$$\hat{g}_\gamma^F = \int_{\tilde{I}} \tilde{\chi}_\gamma(\tilde{s}) g(\mathcal{G}^F(\tilde{s})) |J^F(\tilde{s})| d\tilde{s}.$$

We then apply a n_q-point numerical quadrature defined by $\{\tilde{\xi}_q, \tilde{\rho}_q\}_{q=1}^{n_q}$ to "evaluate" the integral:

$$\hat{g}_\gamma^F \text{“=”} \sum_{q=1}^{n_q} \tilde{\rho}_q \tilde{\chi}_\gamma(\tilde{\xi}_q) g(\mathcal{G}^F(\tilde{\xi}_q)) |J^F(\tilde{\xi}_q)|.$$

We finally map the vector $\hat{g}^F \in \mathbb{R}^{n_s^F}$ to $\hat{f}^K \in \mathbb{R}^{n_s}$ according to

$$\hat{f}_\alpha^K = \hat{g}_\gamma^F, \quad \gamma = 1,\ldots,n_s^F,$$

for $K = \theta_{F\text{-}K}(F)$, $\alpha = \theta_{\tilde{F}\text{-}n}(i,\gamma)$, and $i = \theta_{F\text{-}\tilde{F}}(F)$. This mapping appeals to the relationship (4.27).

4.7 Essential boundary conditions

4.7.1 Homogeneous Dirichlet boundary condition

We recall that Dirichlet boundary conditions are essential boundary conditions, which are explicitly enforced through the choice of the space. As discussed in Section 3.4, we can construct an approximation space $\mathcal{V}_h$ that incorporates the essential boundary condition by first constructing the space $H_h^1(\Omega)$ that does *not* incorporate the essential boundary condition and then removing those nodal shape functions that lie on the Dirichlet boundary $\bar{\Gamma}_D$. We now discuss a convenient implementation of this procedure.

To begin, we first introduce m nodal shape functions for the space $H_h^1(\Omega)$, $\{\bar{\phi}_i\}_{i=1}^m$; the shape functions are associated with the m nodes of the mesh. We then divide the nodes of the mesh into two groups:

$$S_{\Gamma_D} = \{i \in \{1,\ldots,m\} \mid z_i \in \bar{\Gamma}_D\},$$
$$S_{\mathcal{V}_h} = \{1,\ldots,m\} \setminus S_{\Gamma_D}.$$

In words, S_{Γ_D} is a set of nodes that lie on $\bar{\Gamma}_D$, and $S_{\mathcal{V}_H}$ is all other nodes. We denote the cardinality of the two sets by $|\Gamma_D| = n_{\Gamma_D}$ and $|S_{\mathcal{V}_D}| = n$. We next note that the space $\mathcal{V}_h$ that incorporates the homogeneous Dirichlet boundary condition is given by

$$\mathcal{V}_h = \{v \in H_h^1(\Omega) \mid v(z_i) = 0, \forall z_i \in \bar{\Gamma}_D\} = \text{span}\{\bar{\phi}_i \mid i \in S_{\mathcal{V}_h}\}.$$

In words, we obtain the space $\mathcal{V}_h$ by removing basis functions of $H_h^1(\Omega)$ associated with nodes on $\bar{\Gamma}_D$. Note that the dimension of $H_h^1(\Omega)$ is m, the number of Dirichlet boundary nodes is $|S_{\Gamma_D}| = n_{\Gamma_D}$, and the dimension of $\mathcal{V}_h$ is $n = m - n_{\Gamma_D} = |S_{\mathcal{V}_D}|$.

We can construct the stiffness matrix and load vector for $\mathcal{V}_h$ using the above relationship between $H_h^1(\Omega)$ and $\mathcal{V}_h$. We first construct the stiffness matrix and the load vector associated with the $H_h^1(\Omega)$ space $\hat{\bar{A}}_h \in \mathbb{R}^{m \times m}$ and $\hat{\bar{f}}_h \in \mathbb{R}^m$

$$\begin{aligned}
\hat{\bar{A}}_{h,ij} &\equiv a(\bar{\phi}_j, \bar{\phi}_i), \quad i, j = 1, \ldots, m, \\
\hat{\bar{f}}_{h,i} &\equiv \ell(\bar{\phi}_i), \quad i = 1, \ldots, m;
\end{aligned}$$

in practice the construction is carried out using the assembly procedure outlined in Section 4.5. We then remove rows and columns of $\hat{\bar{A}}_h \in \mathbb{R}^{m \times m}$ associated with the Dirichlet-boundary index set S_{Γ_D} to create the stiffness matrix $\hat{A}_h \in \mathbb{R}^{n \times n}$. Or, equivalently, we keep the rows and columns of $\hat{\bar{A}}_h \in \mathbb{R}^{m \times m}$ that are in $S_{\mathcal{V}_h}$; i.e.,

$$\hat{A}_h = \{a(\bar{\phi}_j, \bar{\phi}_i) \mid i, j \in S_{\mathcal{V}_h}\} \in \mathbb{R}^{n \times n}. \tag{4.28}$$

We similarly remove the rows of $\hat{\bar{f}}_h \in \mathbb{R}^m$ associated with the Dirichlet-boundary index set S_{Γ_D} to create the load vector $\hat{f}_h \in \mathbb{R}^n$. Or, equivalently, we keep the rows of $\hat{\bar{f}}_h \in \mathbb{R}^m$ that are in $S_{\mathcal{V}_h}$; i.e.,

$$\hat{f}_h = \{\ell(\bar{\phi}_i) \mid i \in S_{\mathcal{V}_h}\} \in \mathbb{R}^n. \tag{4.29}$$

We finally solve the linear system

$$\hat{A}_h \hat{u}_h = \hat{f}_h$$

for $\hat{u}_h \in \mathbb{R}^n$, which are the coefficients of the basis for $\mathcal{V}_h$.

For a concrete example, consider the $\mathbb{P}^1$ approximation space associated with the mesh shown in Figure 4.17. We recognize that

$$\begin{aligned}
S_{\Gamma_D} &= \{4, 7, 8, 9\} \\
S_{\mathcal{V}_h} &= \{1, 2, 3, 5, 6\}.
\end{aligned}$$

The dimension of $H_h^1(\Omega)$ is $m = 9$, and the associated stiffness matrix is $\hat{\bar{A}}_h \in \mathbb{R}^{9 \times 9}$. The dimension of $\mathcal{V}_h$, which respects the homogeneous Dirichlet boundary condition, is $n = 5$, and the associated stiffness matrix $\hat{A}_h \in \mathbb{R}^{5 \times 5}$ is obtained by eliminating rows and columns in $S_{\Gamma_D} = \{4, 7, 8, 9\}$ from $\hat{\bar{A}}_h \in \mathbb{R}^{9 \times 9}$. Similarly, the load vector $\hat{f}_h \in \mathbb{R}^5$ associated with $\mathcal{V}_h$ is obtained by eliminating rows in $S_{\Gamma_D} = \{4, 7, 8, 9\}$ from the load vector $\hat{\bar{f}}_h \in \mathbb{R}^9$ associated with $H_h^1(\Omega)$.

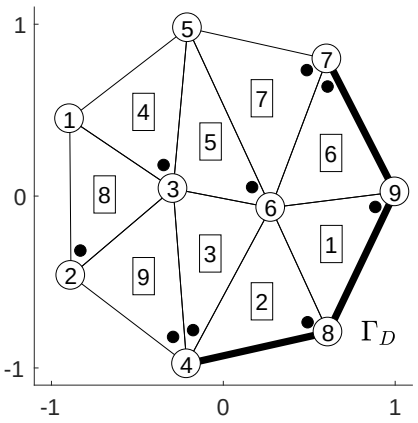


Figure 4.17: A triangulated domain with a Dirichlet boundary.

4.7.2 Nonhomogeneous Dirichlet boundary condition

We now consider an nonhomogeneous Dirichlet boundary condition. We recall that we can think of the trial space with an nonhomogeneous Dirichlet boundary condition as $\mathcal{V}_h^E \equiv u_h^E + \mathcal{V}_h$ where u_h^E is any function in $H_h^1(\Omega)$ such that $u_h^E(z_i) = u^B(z_i)$, $z_i \in \bar{\Gamma}_D$. Our finite element problem is as follows: find $\tilde{u}_h \in \mathcal{V}_h$ such that

$$a(\tilde{u}_h, v) = \ell(v) - a(u_h^E, v) \quad \forall v \in \mathcal{V}_h,$$

then set $u_h = u_h^E + \tilde{u}_h$. For a finite element implementation, it is convenient to choose $u_h^E \in H_h^1(\Omega)$ such that

$$u_h^E(z_i) = \begin{cases} u^B(z_i), & z_i \in \bar{\Gamma}_D, \\ 0, & \text{otherwise.} \end{cases}$$

To obtain an equivalent statement for the coefficient $\hat{\tilde{u}}_h \in \mathbb{R}^n$, we introduce the following matrices and vectors:

- $\hat{A}_h \in \mathbb{R}^{n \times n}$. The stiffness matrix associated with the space $\mathcal{V}_h$ (with the homogeneous Dirichlet boundary condition) as defined in (4.28).
- $\hat{f}_h \in \mathbb{R}^n$. The load vector associated with the space $\mathcal{V}_h$ as defined in (4.29).
- $\hat{u}_h^E \in \mathbb{R}^{n_{\Gamma_D}}$. The coefficients associated with $u_h^E \in H_h^1(\Omega)$ evaluated on Γ_D. We simply set $\hat{u}_h^E = \{u^B(z_i) \mid z_i \in \bar{\Gamma}_D\}$, the boundary function u^B evaluated at the Dirichlet nodes.
- $\hat{B}_h \in \mathbb{R}^{n \times n_{\Gamma_D}}$. The submatrix of the stiffness matrix $\hat{\tilde{A}}_h \in \mathbb{R}^{m \times m}$ associated with the rows in $S_{\mathcal{V}_h}$ and the columns in S_{Γ_D}; i.e., $\hat{B}_h\{a(\bar{\phi}_j, \bar{\phi}_i) \mid i \in S_{\mathcal{V}_h},\ j \in S_{\Gamma_D}\}$.

Then, the linear-algebraic form of the equation for $\hat{\tilde{u}}_h \in \mathbb{R}^n$ is

$$\hat{A}_h \hat{\tilde{u}}_h = \hat{f}_h - \hat{B}_h \hat{u}_h^E.$$

The solution vector $\hat{u}_h \in \mathbb{R}^m$ associated with $u_h \in u_h^E + \mathcal{V}_h$ is then given by setting for the Dirichlet boundary nodes $\{\hat{u}_{h,i}\}_{i \in S_{\Gamma_D}} = \{u^B(z_i)\}_{i \in S_{\Gamma_D}}$ and for all other nodes $\{\hat{u}_{h,i}\}_{i \in S_{\mathcal{V}_h}} = \hat{\tilde{u}}_h$.

4.8 Efficient implementation by BLAS

Before we conclude this lecture, we make a few remarks about an efficient implementation of finite element method on modern computers. While modern computers can carry out billions of floating point operations per second, modern computers also have a deep memory hierarchy and hence not all operations can be carried out as efficiently as others. One way to achieve a high level of computational utilization on a modern computer is to implement many of the operations as linear algebra operations and to use the BLAS (basic linear algebra subroutines) to carry out these operations. BLAS routines are optimized for the particular architecture. There are three different levels of BLAS: BLAS1 deals with vector-vector operations; BLAS2 deals with matrix-vector operations; and BLAS3 deals with matrix-matrix operations. The use of BLAS3 routines, with the most favorable compute-to-memory ratio, is a key to achieve good computational utilization on modern computers. Many of the operations described in this lecture can be written as BLAS operations.

For instance, suppose we want to evaluate the nodal shape functions at quadrature points of a reference element. We first write a routine that evaluates monomials: given n_{pt} points described by a matrix $X \in \mathbb{R}^{n_{\text{pt}} \times d}$, evaluate the matrix $\Psi(X) \in \mathbb{R}^{n_{\text{pt}} \times n_s}$, where $(\Psi(X))_{ij}$ is the j-th monomial basis function evaluated at the i-th point. Using this routine, we can express the Vandermonde matrix as

$$V = \Psi(X_{\text{int}}) \in \mathbb{R}^{n_s \times n_s},$$

where $X_{\text{int}} \in \mathbb{R}^{n_s \times d}$ is the set of Lagrange interpolation nodes. To compute the nodal shape functions at quadrature points $X_{\text{quad}} \in \mathbb{R}^{n_{\text{quad}} \times d}$, we appeal to the fact that the Vandermonde matrix is the inverse of the nodal basis coefficient matrix and invoke

$$\Phi(X_{\text{quad}}) = \Psi(X_{\text{quad}})V^{-1} \in \mathbb{R}^{n_{\text{quad}} \times n_s}.$$

(In the actual implementation, the inverse should never be explicitly computed but its action should be computed through a linear solve; in MATLAB, the above can be computed using the "forward slash" operator as $\Phi(X_{\text{quad}}) = \Psi(X_{\text{quad}})/V$.) This is an efficient and concise way to compute nodal shape functions using matrix operations.

As another example, suppose we now wish to evaluate the mass matrix on the reference element, $\tilde{M} \equiv \mathbb{R}^{n_s \times n_s}$ such that

$$\tilde{M}_{ij} = \int_{\tilde{K}} \tilde{\phi}_i(\tilde{x})\tilde{\phi}_j(\tilde{x})d\tilde{x}, \quad i,j = 1,\ldots,n_s.$$

If the shape functions are polynomials of degree p, then we can exactly evaluate the mass matrix by using a quadrature rule of degree $2p$:

$$\tilde{M}_{ij} = \sum_{q=1}^{n_q} \tilde{\rho}_q \tilde{\phi}_i(\tilde{\xi}_q)\tilde{\phi}_j(\tilde{\xi}_q), \quad i,j = 1,\ldots,n_s.$$

Now, we can combine (i) a vector of quadrature weights $\tilde{\rho} \in \mathbb{R}^{n_{\text{quad}}}$ and (ii) shape functions evaluated at quadrature points $\Phi(X_{\text{quad}}) \in \mathbb{R}^{n_{\text{quad}} \times n_s}$ to evaluate the mass matrix as

$$\tilde{M} = \Phi(X_{\text{quad}})^T \text{diag}(\tilde{\rho})\Phi(X_{\text{quad}}).$$

(In MATLAB, it is more efficient to replace $\text{diag}(\tilde{\rho})\Phi$ with a binary singleton expansion operator bsxfun(@times, $\tilde{\rho}, \Phi$).) The mass matrix can then be efficiently computed using a BLAS3 routine.

With some planning, a significant fraction — or more precisely most computationally intense parts — of finite element code can be expressed in terms of BLAS routines. The use of BLAS is important for both compiled languages (e.g., C, C++) and interpreted languages (e.g., MATLAB, Python). However, it is arguably more important for the interpreted languages because the efficiency of interpreted languages can be quite limited for simple but computationally intense operations.

4.9 Summary

We summarize key points of this lecture:

1. Given a polynomial space and a set of nodes, the associated nodal shape functions $\{\tilde{\phi}_\alpha\}$ can be identified using the Vandermonde method.

2. A finite element is formally defined by (i) a domain, (ii) a finite-dimensional approximation space, and (iii) degrees of freedom used to describe functions in the space.

3. A reference element $\tilde{K}$ is mapped to a physical element K by using a polynomial geometry mapping $\mathcal{G}^K$. The differentiation of the mapping yields the Jacobian J^K, from which we compute the determinant of the Jacobian $|J^K|$ and the inverse Jacobian $(J^K)^{-1}$.

4. In $\mathbb{R}^2$, a reference line segment $\tilde{I}$ is mapped to a physical facet F by using a polynomial geometry mapping $\mathcal{G}^F$. The differentiation of the mapping yields the Jacobian J^F, from which we compute the determinant of the Jacobian J^F and the outward-pointing normal vector n^F.

5. Physical shape functions on a given element, $\{\phi_\alpha^K\}$, are obtained by mapping the reference shape functions $\{\tilde{\phi}_\alpha\}$ through the geometry mapping $\mathcal{G}^K$.

6. On a reference line segment, a n_q-point Gauss quadrature rule exactly integrates polynomials of degree up to $2n_q - 1$. The rule is most efficient in the sense that it requires the fewest number of points to integrate a polynomial of a given degree.

7. Efficient quadrature rules for canonical domains in higher dimensions also exist; however, they are not as well standardized as Gauss quadrature in one dimension.

8. To evaluate the domain-integration terms in local stiffness matrices and load vectors, we (i) map the integration domain from K to $\tilde{K}$ using the geometry mapping $\mathcal{G}^K$, (ii) evaluate the integrand, and in particular the physical basis functions $\{\phi_\alpha^K\}$ and their gradients, and (iii) invoke an appropriate quadrature rule. A similar procedure exists for terms that require boundary integration.

9. Essential boundary conditions are imposed by eliminating shape functions on $\bar{\Gamma}_D$ from the approximation space. The task translates to the elimination of the rows and columns associated with $\bar{\Gamma}_D$ from the (global) stiffness matrix and load vector.

10. An efficient finite element solver can be implemented on a modern computer using BLAS routines.

Lecture 5

Polynomial interpolation in Sobolev spaces

5.1 Motivation

As we have seen in the previous lectures, the finite element method approximates the solution to the variational problem in a finite-dimensional approximation space. In the finite element method based on h-refinement, we consider a sequence of piecewise polynomial spaces of various characteristic element diameter h. The accuracy of a given finite element approximation depends on the ability of the underlying piecewise polynomial space to approximate the solution. In this lecture, we characterize the error associated with piecewise polynomial interpolations of functions in Sobolev spaces.

For the majority of the lecture, we consider linear polynomial interpolation in $\mathbb{R}^1$; the simplified setting allows us to analyze interpolation errors without introducing technical tools that are beyond the scope of this course while still capturing the essence of interpolation error theory. In Section 5.7, we introduce, without proofs, more general results for higher-degree polynomial interpolation in $\mathbb{R}^{d\geq 1}$.

5.2 Linear interpolation error for C^2 functions in $\mathbb{R}^1$

In this section, we analyze the error associated with the piecewise linear interpolation of C^2 functions in one dimension. To this end, we introduce a domain $\Omega \equiv [a,b] \subset \mathbb{R}^1$ and interpolation nodes $a \equiv z_1 < \cdots < z_n \equiv b$. Given a function w, its piecewise linear polynomial interpolant $\mathcal{I}_h w$ is a piecewise linear polynomial,

$$(\mathcal{I}_h w)|_{[z_i,z_{i+1}]} \in \mathbb{P}^1([z_i,z_{i+1}]), \quad i = 1,\ldots,n-1,$$

that satisfies the interpolation conditions,

$$(\mathcal{I}_h w)(z_i) = w(z_i), \quad i = 1,\ldots,n.$$

Assuming $w \in C^0([a,b])$, the piecewise polynomial interpolant exists and is unique. The subscript h on $\mathcal{I}_h w$ emphasizes that the interpolant depends on the node spacing h.

In the context of finite element analysis, our goal is to establish interpolation error bounds for functions in Sobolev spaces $H^k(\Omega)$, the space of functions whose *weak* derivatives of order up to k are square integrable (in the Lebesgue sense). But, we first digress and provide a more "classical" interpolation error bounds for functions in $C^k(\Omega)$, the space of functions whose (strong) derivatives of order up to k are continuous (in the pointwise sense).

We first analyze the error for a single-segment interpolant $\mathcal{I}_h w$ over $[a,b]$.

Proposition 5.1. Let $w \in C^2([a,b])$ and $\mathcal{I}_h w \in \mathbb{P}^1([a,b])$ be the (single-segment) linear interpolant. Then, the interpolation error is bounded by

$$|w(x) - (\mathcal{I}_h w)(x)| \leq \frac{1}{8} \max_{s\in[a,b]} |w''(s)|(b-a)^2 \quad \forall x \in [a,b].$$

Proof. We first introduce an auxiliary function

$$g(s) = w(s) - (\mathcal{I}_h w)(s) - \left(\frac{w(x) - (\mathcal{I}_h w)(x)}{(x-a)(x-b)}\right)(s-a)(s-b).$$

We note that $g(x) = 0$ by construction, and $g(a) = g(b) = 0$ because $\mathcal{I}_h w$ interpolants w at the endpoints. Hence, g has at least three roots in $[a,b]$. By Rolle's theorem, g' has at least two root in $[a,b]$. Invoking the Rolle's theorem again, g'' has at least one root in $[a,b]$. Let ξ be one of these roots: $g''(\xi) = 0$. We now compute g'':

$$g''(s) = w''(s) - \left(\frac{w(x) - (\mathcal{I}_h w)(x)}{(x-a)(x-b)}\right) \cdot 2.$$

Note that $(\mathcal{I}_h w)'' = 0$ since $\mathcal{I}_h w$ is a linear function. We now evaluate the expression at ξ to obtain

$$0 = w''(\xi) - \left(\frac{w(x) - (\mathcal{I}_h w)(x)}{(x-a)(x-b)}\right) \cdot 2,$$

or, equivalently,

$$w(x) - (\mathcal{I}_h w)(x) = \frac{1}{2} w''(\xi)(x-a)(x-b).$$

We finally note that $|w''(\xi)| \leq \max_{s\in[a,b]} |w''(s)|$ and $|(x-a)(x-b)| \leq (b-a)^2/4$. □

Proposition 5.2. Let $\bar{\Omega} \subset \mathbb{R}^1$ and $\mathcal{I}_h w$ be the piecewise linear interpolant associated with a triangulation $\mathcal{T}_h \equiv \{K\}$ with $h \equiv \max_{K\in\mathcal{T}_h} |K|$. If $w \in C^0(\bar{\Omega}) \cap C^2(\mathcal{T}_h)$, then the interpolation error is bounded by

$$|w(x) - (\mathcal{I}_h w)(x)| \leq \frac{1}{8} \max_{K\in\mathcal{T}_h} \max_{s\in K} |w''(s)|h^2 \quad \forall x \in \bar{\Omega},$$

where $C^2(\mathcal{T}_h) \equiv \oplus_{K\in\mathcal{T}_h} C^2(\bar{K})$, the space of piecewise C^2 functions.

Proof. Apply Proposition 5.1 to each segment of the piecewise linear interpolant. □

Proposition 5.2 shows that the maximum error in the piecewise linear interpolation is a function of (i) the maximum second derivative $\max_{K\in\mathcal{T}_h}\max_{s\in K}|w''(s)|$ in the "broken" space and (ii) the node spacing h. Note that we require the underlying function to be only *piecewise* $C^2(\mathcal{T}_h)$ continuous, instead of *global* $C^2(\Omega)$ continuous, because the interpolant is constructed independently for each segment. While this "classical" interpolation error bound is useful in many scenarios that involve functions in C^k spaces — with continuous (strong) derivatives in the pointwise sense —, it is not the natural choice for finite element analysis that involves functions in H^k Sobolev spaces — with square integrable weak derivatives in the Lebesgue sense. In the following sections, we introduce interpolation error bounds for functions in H^k Sobolev spaces.

5.3 Preliminary: Rayleigh quotient

By way of preliminary, we first introduce a technical tool required to analyze interpolation error in Sobolev spaces: the *Rayleigh quotient* and the associated bounds. We first introduce Rayleigh quotients for linear operators in $\mathbb{R}^n$; i.e., the matrices $\mathbb{R}^{n\times n}$.

Definition 5.3 (Rayleigh quotient (matrices))**.** Let $A\in\mathbb{R}^{n\times n}$ be a symmetric matrix. The associated Rayleigh quotient is $R_A:\mathbb{R}^n\to\mathbb{R}$ such that

$$R_A(x)\equiv\frac{x^TAx}{x^Tx}.$$

Proposition 5.4 (Bound of Rayleigh quotient (matrices))**.** Let $(w_k,\lambda_k)\in\mathbb{R}^n\times\mathbb{R}$, $k=1,\ldots,n$, be the eigenpairs of a symmetric matrix $A\in\mathbb{R}^{n\times n}$. Then, the Rayleigh quotient of A is bounded by

$$\min_k\lambda_k\le R_A(x)\le\max_k\lambda_k\quad\forall x\in\mathbb{R}^n.$$

Proof. Because A is symmetric, there exists a set of eigenvectors $\{w_k\}_{k=1}^n$ that forms an orthonormal basis of $\mathbb{R}^n$. Hence, $\forall x\in\mathbb{R}^n$, $\exists\hat{x}\in\mathbb{R}^n$ such that $x=\sum_{k=1}^n\hat{x}_kw_k$. The Rayleigh quotient can then be expressed as an weighted sum of eigenvalues

$$R_A(x=\sum_{k=1}^n\hat{x}_kw_k)\equiv\frac{\sum_{k=1}^n\lambda_k\hat{x}_k^2}{\sum_{k=1}^n\hat{x}_k^2}.$$

The minimum value of the weighted sum is obtained for $\hat{x}_k=e_1\in\mathbb{R}^n$, a vector with 1 in the first entry and zero elsewhere, and the associated minimum is λ_1. Similarly, the maximum value of the weighted sum is obtained for $\hat{x}_k=e_n\in\mathbb{R}^n$, a vector with 1 in the last entry and zero elsewhere, and the associated maximum is λ_n. □

We now generalize Rayleigh quotient to bilinear forms in Hilbert spaces.

Definition 5.5 (Rayleigh quotient (Hilbert spaces))**.** Let $\mathcal{V}$ be a Hilbert space endowed with an inner product $(\cdot,\cdot)_\mathcal{V}:\mathcal{V}\times\mathcal{V}\to\mathbb{R}$, and $a:\mathcal{V}\times\mathcal{V}\to\mathbb{R}$ be a symmetric, positive bilinear form. The associated Rayleigh quotient is $R_a:\mathcal{V}\to\mathbb{R}$ such that

$$R_a(v)\equiv\frac{a(v,v)}{(v,v)_\mathcal{V}}.$$

Proposition 5.6 (Bounds of Rayleigh quotient (Hilbert spaces))**.** Let $\mathcal{V}$ be a Hilbert space endowed with an inner product $(\cdot,\cdot)_{\mathcal{V}} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, and $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a symmetric, positive bilinear form. Consider also the following eigenproblem: find $(w_k, \lambda_k) \in \mathcal{V} \times \mathbb{R}$, $k \in \mathbb{N}$, such that $\|w_k\|_{\mathcal{V}} = 1$ and

$$a(w_k, v) = \lambda_k (w_k, v)_{\mathcal{V}} \quad \forall v \in \mathcal{V}.$$

Then, the Rayleigh quotient of $a(\cdot,\cdot)$ is bounded by

$$\inf\{\lambda_k\} \le R_a(v) \le \sup\{\lambda_k\}.$$

Moreover, if $\inf\{\lambda_k\} > 0$, then

$$\inf\{\lambda_k\} = \alpha,$$

where $\alpha > 0$ is the $\mathcal{V}$-coercivity constant; if $\sup\{\lambda_k\} < \infty$, then

$$\sup\{\lambda_k\} = \gamma,$$

where $\gamma < \infty$ is the $\mathcal{V}$-continuity constant.

5.4 The $L^2(\Omega)$ error of linear interpolant of $H^2(\mathcal{T}_h)$ functions in $\mathbb{R}^1$

We now consider piecewise linear interpolation for $H^2(\Omega)$ functions in $\mathbb{R}^1$. (More precisely, we consider functions in a broken space $H^2(\mathcal{T}_h) \supset H^2(\Omega)$, which will be introduced shortly.) We first provide a definition of the interpolant.

Definition 5.7 (piecewise linear interpolant in $\mathbb{R}^1$)**.** Let $\Omega \subset \mathbb{R}^1$. Consider a triangulation $\mathcal{T}_h \equiv \{K_i\}_{i=1}^{n_e}$ delineated by $n = n_e + 1$ nodes $\{z_i\}_{i=1}^n$ such that $K_i \equiv (z_i, z_{i+1})$ and $h \equiv \max_{K \in \mathcal{T}_h} |K|$. Consider the associated approximation space

$$\mathcal{V}_h \equiv \{v \in H^1(\Omega) \mid v|_K \in \mathbb{P}^1(K),\ \forall K \in \mathcal{T}_h\}.$$

For $w \in H^1(\Omega)$, the interpolant $\mathcal{I}_h w$ is a unique member of $\mathcal{V}_h$ that satisfies the interpolation condition

$$(\mathcal{I}_h w)(z_i) = w(z_i), \quad i = 1, \dots, n.$$

Note that the construction of the interpolant is straightforward given a Lagrange basis $\{\phi_i\}_{i=1}^n$ of $\mathcal{V}_h$: $\mathcal{I}_h w = \sum_{i=1}^n w(z_i)\phi_i$.

We now wish to bound the L^2 norm of the interpolation error. We take a three-step strategy: (i) we first introduce an embedding constant between two relevant spaces; (ii) we next derive an interpolation error bound for a unit line segment $\tilde{I} \equiv (a, b)$; (iii) we finally make a homogeneity argument to establish an interpolation error bound for piecewise linear interpolant.

Lemma 5.8 ($L^2(\tilde{I})$-$H_0^2(\tilde{I})$ embedding constant)**.** Let $\tilde{I} \equiv (0, 1)$ and $H_0^2(\tilde{I}) \equiv \{v \in H^2(\tilde{I}) \mid v(x = 0) = v(x = 1) = 0\}$. Then

$$\rho_{L^2(\tilde{I})\text{-}H_0^2(\tilde{I})} \equiv \sup_{v \in H_0^2(\tilde{I})} \frac{\|v\|_{L^2(\tilde{I})}}{|v|_{H^2(\tilde{I})}} = \frac{1}{\pi^2}$$

Proof. We first introduce a Rayleigh quotient

$$R(v) = \frac{\|v\|^2_{L^2(\tilde{I})}}{|v|^2_{H^2(\tilde{I})}} = \frac{\int_{\tilde{I}} v^2 dx}{\int_{\tilde{I}} (v'')^2 dx}.$$

The associated eigenproblem is as follows: find eigenpairs $(u_k, \lambda_k) \in H^2_0(\tilde{I}) \times \mathbb{R}$, $k \in \mathbb{Z}_{>0}$, such that

$$\int_{\tilde{I}} v u_k dx = \lambda_k \int_{\tilde{I}} v'' u_k'' dx \quad \forall v \in H^2_0(\tilde{I}).$$

To identify the strong form of the eigenproblem, we integrate by parts twice the right hand side to obtain

$$\begin{aligned}\int_{\tilde{I}} v u_k dx &= \lambda_k \left(- \int_{\tilde{I}} v u_k''' dx + [v' u_k'']_{x=0}^1 \right) \\ &= \lambda_k \left(\int_{\tilde{I}} v u_k'''' dx + [v' u_k'']_{x=0}^1 - [v u_k''']_{x=0}^1 \right) \quad \forall v \in H^2_0(\tilde{I}).\end{aligned}$$

We recognize $v(x=0) = v(x=1) = 0$ and rearrange the expression to obtain

$$\int_I v(u_k - \lambda_k u_k'''') dx - \lambda_k [v' u_k'']_{x=0}^1 = 0 \quad \forall v \in H^2_0(\tilde{I}).$$

We recognize that the strong form of the eigenproblem is

$$u_k = \lambda_k u_k'''' \quad \text{in } \tilde{I}$$

with boundary conditions

$$u_k(x=0) = u_k(x=1) = u_k''(x=0) = u_k''(x=1) = 0.$$

The eigenpairs are

$$\begin{aligned}u_k &= \sin(k\pi x), \\ \lambda_k &= \frac{1}{k^4 \pi^4}, \quad k \in \mathbb{N}.\end{aligned}$$

The upper bound of the Rayleigh quotient, which is equal to $\rho^2_{L^2(\tilde{I})\text{-}H^2_0(\tilde{I})}$, is given for $k = 1$ and is $1/\pi^4$. Hence the embedding constant is $\rho_{L^2(\tilde{I})\text{-}H^2_0(\tilde{I})} = 1/\pi^2$. □

Proposition 5.9 (linear interpolation error on $\tilde{I} \equiv (0,1)$)**.** Let $\tilde{I} \equiv (0,1)$. If $\tilde{w} \in H^2(\tilde{I})$, then the linear interpolation error is bounded by

$$\|\tilde{w} - \mathcal{I}\tilde{w}\|_{L^2(\tilde{I})} \leq \frac{1}{\pi^2} |\tilde{w}|_{H^2(\tilde{I})},$$

and this bound is sharp.

Proof. For any $\tilde{w} \in H^2(\tilde{I})$, $\tilde{w} - \mathcal{I}\tilde{w} \in H_0^2(\tilde{I})$ because the interpolant $\mathcal{I}\tilde{w} \in H^2(\tilde{I})$ matches the function w at the endpoints. It hence follows that

$$\begin{aligned}
\|\tilde{w} - \mathcal{I}\tilde{w}\|_{L^2(\tilde{I})} &= \frac{\|\tilde{w} - \mathcal{I}\tilde{w}\|_{L^2(\tilde{I})}}{|\tilde{w} - \mathcal{I}\tilde{w}|_{H^2(\tilde{I})}} |\tilde{w} - \mathcal{I}\tilde{w}|_{H^2(\tilde{I})} \\
&\leq \sup_{v \in H_0^2(\Omega)} \frac{\|v\|_{L^2(\tilde{I})}}{|v|_{H^2(\tilde{I})}} |\tilde{w} - \mathcal{I}\tilde{w}|_{H^2(\tilde{I})} \qquad \text{(maximization of the ratio)} \\
&= \sup_{v \in H_0^2(\Omega)} \frac{\|v\|_{L^2(\tilde{I})}}{|v|_{H^2(\tilde{I})}} |\tilde{w}|_{H^2(\tilde{I})} \qquad ((\mathcal{I}\tilde{w})'' = 0 \text{ since } \mathcal{I}\tilde{w} \in \mathbb{P}^1(\tilde{I})) \\
&= \frac{1}{\pi^2} |\tilde{w}|_{H^2(\tilde{I})}, \qquad \text{(Lemma 5.8)}
\end{aligned}$$

which is the desired bound. □

We now define *broken* Sobolev spaces suitable for the analysis of piecewise polynomial interpolants.

Definition 5.10 (broken Sobolev space $H^k(\mathcal{T}_h)$)**.** Consider $\Omega \subset \mathbb{R}^d$ and an associated triangulation $\mathcal{T}_h \equiv \{K_i\}_{i=1}^{n_e}$ comprises n_e (open) elements such that (i) $K_i \cap K_j = \emptyset$, $i \neq j$, and (ii) $\cup_{K \in \mathcal{T}_h} \bar{K} = \bar{\Omega}$. The space $H^k(\mathcal{T}_h)$ is endowed with an inner product

$$(w, v)_{H^k(\mathcal{T}_h)} \equiv \sum_{K \in \mathcal{T}_h} (w, v)_{H^k(K)},$$

the associated induced norm $\|w\|_{H^k(\mathcal{T}_h)} \equiv \sqrt{(w, w)_{H^k(\mathcal{T}_h)}}$, and comprises functions

$$H^k(\mathcal{T}_h) \equiv \{v \mid \|v\|_{H^k(\mathcal{T}_h)} < \infty\}.$$

We also introduce the associated semi-norm

$$|w|^2_{H^k(\mathcal{T}_h)} \equiv \sum_{K \in \mathcal{T}_h} |w|^2_{H^k(K)}.$$

We finally make a *homogeneity (or scaling) argument* to obtain an interpolation error bound for piecewise linear interpolants.

Proposition 5.11 (piecewise linear interpolation error in $\mathbb{R}^1$ (L^2))**.** Let $\mathcal{I}_h w$ be the piecewise linear interpolant in Definition 5.7. If $w \in H^1(\Omega) \cap H^2(\mathcal{T}_h)$, then

$$\|w - \mathcal{I}_h w\|_{L^2(\Omega)} \leq \frac{h^2}{\pi^2} |w|_{H^2(\mathcal{T}_h)}.$$

Proof. We first consider the interpolation error for a single element $K \in \mathcal{T}_h$ of length h. To this end, we map the function $w|_K$ on $K \equiv (z_i, z_i + h)$ to the function $\tilde{w}$ on $\tilde{I} \equiv (0, 1)$. We associate a point $x \in K$ with $\tilde{x} \in \tilde{K}$ according to $x = z_i + h\tilde{x}$; the functions are related by

$$\tilde{w}(\tilde{x}) = w(x \equiv z_i + h\tilde{x}) \quad \forall \tilde{x} \in \tilde{I}.$$

By the chain rule, the second derivatives are related by

$$\tilde{w}''(\tilde{x}) = w''(x)h^2$$

It hence follows that

$$\begin{aligned}\|w - \mathcal{I}_h w\|^2_{L^2(K)} &\equiv \int_K (w(x) - (\mathcal{I}_h w)(x))^2 dx = \int_{\tilde{I}} (\tilde{w}(\tilde{x}) - (\mathcal{I}\tilde{w})(\tilde{x}))^2 h d\tilde{x} \\ &\leq \frac{1}{\pi^4}\int_{\tilde{I}} \tilde{w}''(\tilde{x})^2 h d\tilde{x} = \frac{1}{\pi^4}\int_K (w''(x)h^2)^2 h h^{-1} dx = \frac{h^4}{\pi^4}|w|^2_{H^2(K)}.\end{aligned}$$

We now sum over of the elements in $\mathcal{T}_h$ to obtain

$$\|w - \mathcal{I}_h w\|^2_{L^2(\Omega)} = \sum_{K\in\mathcal{T}_h} \|w - \mathcal{I}_h w\|^2_{L^2(K)} \leq \sum_{K\in\mathcal{T}_h} \frac{h^4}{\pi^4}|w|^2_{H^2(K)} = \frac{h^4}{\pi^4}|w|^2_{H^2(\mathcal{T}_h)}.$$

Taking the square root of the equation yields the desired bound. □

Proposition 5.11 shows that the $L^2(\Omega)$ error associated with the piecewise linear interpolation of $H^2(\Omega)$ functions (i) depends on the $H^2(\mathcal{T}_h)$ semi-norm of the underlying function and (ii) converges as h^2. While the result is similar to Proposition 5.2 for C^2 functions, the regularity requirement for the underlying function is weaker for Proposition 5.11 for H^2 functions. Indeed, the underlying function need not be twice differentiable in the strong sense; its weak second derivative needs only be square integrable (in the broken space). The fact that the function needs only be in $H^2(\mathcal{T}_h) \supset H^2(\Omega)$ is a direct consequence of the piecewise construction of the interpolant; this relaxation will play an important role for solutions that are only piecewise smooth, which arise in the presence of, for example, discontinuous interior/boundary heat source or discontinuous conductivity for the heat equation.

5.5 The $H^1(\Omega)$ error of linear interpolant of $H^2(\mathcal{T}_h)$ functions in $\mathbb{R}^1$

We now analyze the error in the piecewise linear interpolation of $H^2(\mathcal{T}_h)$ functions in a norm stronger than the $L^2(\Omega)$ norm: the $H^1(\Omega)$ norm. The analysis follows essentially the same argument as that used for the $L^2(\Omega)$ norm of the error in the previous section.

Lemma 5.12 ($H^1(\tilde{I})$-$H^2_0(\tilde{I})$ embedding constant)**.** Let $\tilde{I} \equiv (0,1)$ and $H^2_0(\tilde{I}) \equiv \{v \in H^2(\tilde{I}) \mid v(x = 0) = v(x = 1) = 0\}$. Then

$$\rho_{H^1(\tilde{I})\text{-}H^2_0(\tilde{I})} \equiv \sup_{v\in H^2_0(\Omega)} \frac{|v|_{H^1(\tilde{I})}}{|v|_{H^2(\tilde{I})}} = \frac{1}{\pi}.$$

Proof. Proof is similar to Proposition 5.8 for the $L^2(\tilde{I})$-$H^2_0(\tilde{I})$ embedding constant. Here we provide a sketch. The eigenproblem associated with the Rayleigh quotient is as follows: find eigenpairs $(u_k, \lambda_k) \in H^2_0(\tilde{I}) \times \mathbb{R}$, $k \in \mathbb{Z}_{>0}$, such that

$$\int_{\tilde{I}} v' u_k' dx = \lambda_k \int_{\tilde{I}} v'' u_k'' dx \quad \forall v \in H^2_0(\tilde{I}).$$

We can readily show that the eigenpairs are $u_k = \sin(k\pi x)$ and $\lambda_k = 1/(k^2\pi^2)$. The maximum eigenvalue is $1/\pi^2$, and the embedding constant, which is the square root of the Rayleigh quotient, is bounded by $1/\pi$. □

Proposition 5.13 (piecewise linear interpolation error in $\mathbb{R}^1$ (H^1))**.** Let $\mathcal{I}_h w$ be the piecewise linear interpolant in Definition 5.7. If $w \in H^1(\Omega) \cap H^2(\mathcal{T}_h)$, then

$$|w - \mathcal{I}_h w|_{H^1(\Omega)} \leq \frac{h}{\pi} |w|_{H^2(\mathcal{T}_h)}.$$

Proof. Proof is similar to the analysis for the L^2 error in Proposition 5.11 and uses the homogeneity argument, except now we appeal to Lemma 5.12. We here omit the proof for brevity. □

Corollary 5.14. In the same setting as Proposition 5.13, the $H^1(\Omega)$ norm of the interpolation error is bounded by

$$\|w - \mathcal{I}_h w\|_{H^1(\Omega)} \leq \left(\frac{h^2}{\pi^2} + \frac{h^4}{\pi^4} \right)^{1/2} |w|_{H^2(\mathcal{T}_h)}.$$

For h sufficiently small, $\exists C < \infty$ such that

$$\|w - \mathcal{I}_h w\|_{H^1(\Omega)} \leq Ch |w|_{H^2(\mathcal{T}_h)}.$$

Proposition 5.13 shows that the $H^1(\Omega)$ error associated with the pieceiwse linear interpolation of $H^2(\mathcal{T}_h)$ functions (i) depends on the $H^2(\mathcal{T}_h)$ semi-norm of the underlying function and (ii) converges as h^1. Note that the $H^1(\Omega)$ norm is a stronger norm than the $L^2(\Omega)$ norm, and hence the $H^1(\Omega)$ error converges at a lower rate than the $L^2(\Omega)$ error.

5.6 The $L^2(\Omega)$ error of linear interpolant of $H^1(\Omega)$ functions in $\mathbb{R}^1$

We now consider the case where the underlying function is only in $H^1(\Omega)$ for $\Omega \subset \mathbb{R}^1$. We recall that $H^1(\Omega)$ space includes rather irregular functions; for example, a continuous function with a kink, which would not be in $C^1(\bar{\Omega})$, is in $H^1(\Omega)$.

Proposition 5.15. Let $\tilde{I} \equiv (0,1)$, and $\mathcal{I}v \in \mathbb{P}^1(\tilde{I})$ be a linear interpolant of $v \in H^1(\tilde{I})$ so that $(\mathcal{I}v)(x=0) = v(x=0)$ and $(\mathcal{I}v)(x=1) = v(x=1)$. Then,

$$\rho \equiv \sup_{v \in H^1(\tilde{I})} \frac{\|v - \mathcal{I}v\|_{L^2(\tilde{I})}}{|v|_{H^1(\tilde{I})}} = \frac{1}{\pi}.$$

Proof. We first expand the denominator to obtain

$$|v|^2_{H^1(\tilde{I})} = |\mathcal{I}v + (v - \mathcal{I}v)|^2_{H^1(\tilde{I})} = |\mathcal{I}v|^2_{H^1(\tilde{I})} + |v - \mathcal{I}v|^2_{H^1(\tilde{I})} + 2 \int_{\tilde{I}} (\mathcal{I}v)'(v - \mathcal{I}v)' dx.$$

The last term of the expansion vanishes according to

$$\int_{\tilde{I}} (\mathcal{I}v)'(v - \mathcal{I}v)' dx = - \int_{\tilde{I}} \underbrace{(\mathcal{I}v)''}_{=0 \,:\, \mathcal{I}v \text{ is linear}} (v - \mathcal{I}v) dx + \underbrace{\left[(\mathcal{I}v)'(v - \mathcal{I}v) \right]_{x=0}^{1}}_{=0 \,:\, \text{interpolation condition}} = 0,$$

and hence $|v|^2_{H^1(\tilde{I})} = |\mathcal{I}v|^2_{H^1(\tilde{I})} + |v - \mathcal{I}v|^2_{H^1(\tilde{I})}$. It follows that

$$\rho^2 \equiv \sup_{v \in H^1(\tilde{I})} \frac{\|v - \mathcal{I}v\|^2_{L^2(\tilde{I})}}{|\mathcal{I}v|^2_{H^1(\tilde{I})} + |v - \mathcal{I}v|^2_{H^1(\tilde{I})}} \leq \sup_{v \in H^1(\tilde{I})} \frac{\|v - \mathcal{I}v\|^2_{L^2(\tilde{I})}}{|v - \mathcal{I}v|^2_{H^1(\tilde{I})}} = \sup_{v \in H^1_0(\tilde{I})} \frac{\|v\|^2_{L^2(\tilde{I})}}{|v|^2_{H^1(\tilde{I})}}.$$

The inequality is sharp for any $v \in H_0^1(\tilde{I})$ so that $\mathcal{I}v = 0$; hence

$$\rho^2 = \sup_{v \in H_0^1(\tilde{I})} \frac{\|v\|_{L^2(\tilde{I})}^2}{|v|_{H^1(\tilde{I})}^2}.$$

The constant ρ_K^2 is an Rayleigh quotient whose bound is given by the following eigenproblem: find eigenpairs $(u_k, \lambda_k) \in H_0^1(\tilde{I}) \times \mathbb{R}$, $k = 1, 2, \ldots$, such that

$$\int_{\tilde{I}} v u_k dx = \lambda_k \int_{\tilde{I}} v' u_k' dx \quad \forall v \in H_0^1(\tilde{I}).$$

The eigenpairs are

$$u_k(x) = \sin(k\pi x) \quad \text{and} \quad \lambda_k = \frac{1}{k^2\pi^2}, \quad k = 1, 2, \ldots.$$

The maximum eigenvalue is $1/\pi^2$, and hence $\rho^2 = 1/\pi^2$. □

Proposition 5.16 (piecewise linear interpolation error in $\mathbb{R}^1$ (L^2))**.** Let $\mathcal{I}_h w$ be the piecewise linear interpolant in Definition 5.7. If $w \in H^1(\Omega)$, then

$$\|w - \mathcal{I}_h w\|_{L^2(\Omega)} \leq \frac{h}{\pi} |w|_{H^1(\Omega)},$$

Proof. Proof is similar to the analysis for the L^2 interpolation error of $H^2(\Omega)$ functions in Proposition 5.11 and uses the homogeneity argument, except now we appeal to Lemma 5.15. We here omit the proof for brevity. □

The proposition shows that, if the underlying function is in $H^1(\Omega)$ instead of $H^1(\Omega) \cap H^2(\mathcal{T}_h)$, then the L^2 error of the piecewise linear interpolant converges as h^1 instead of h^2. The convergence rate is reduced due to the limited regularity of the underlying function.

5.7 Generalization: piecewise $\mathbb{P}^p$ interpolation in $\mathbb{R}^d$

The interpolation error bound obtained in Sections 5.4, 5.5, and 5.6 can be generalized to (i) higher dimensions, (ii) higher-degree polynomials, and (iii) non-uniform meshes. To this end, we consider piecewise degree-p polynomial spaces of the form

$$\mathcal{V}_h \equiv \{v \in H^1(\Omega) \mid v|_K \in \mathbb{P}^p(K), \ \forall K \in \mathcal{T}_h\}$$

associated with a triangulation $\mathcal{T}_h \equiv \{K_i\}_{i=1}^{n_e}$ delineated by n nodes $\{z_i\}_{i=1}^n$. Given $w \in C^0(\bar{\Omega})$, the piecewise $\mathbb{P}^p$ interpolant is the function in $\mathcal{V}_h$ that satisfies the interpolation condition

$$(\mathcal{I}_h w)(z_i) = w(z_i), \quad i = 1, \ldots, n.$$

Given a Lagrange basis $\{\phi_i\}_{i=1}^n$ of $\mathcal{V}_h$, we can readily construct the interpolant: $\mathcal{I}_h w = \sum_{i=1}^n w(z_i)\phi_i$.

In order to discuss the convergence of polynomial interpolants in higher dimensions, we first introduce the notation of a *shape-regular* or *non-degenerate* family of triangulations.

Definition 5.17 (shape-regular meshes)**.** A family of meshes $\{\mathcal{T}_h\}_{h>0}$ is said to be shape-regular if there exists $r_0 < \infty$ such that

$$r_K \equiv \frac{h_K}{\rho_K} \leq r_0, \quad \forall K \in \mathcal{T}_h, \quad \forall h,$$

where h_K is the diameter of the element K, and ρ_K is the maximum diameter of the largest ball that can be inscribed in K.

We make a few remarks. First, we note that shape regularity is a property associated with a *family* (or sequence) of triangulations of various h, and not a single triangulation. Second, in one dimension $\rho_K = h_K$, and any triangulation is shape-regular. Third, for triangles, $\frac{h_K}{\rho_K} \leq \frac{2}{\sin(\theta_K)}$, where θ_K is the smallest angle; the triangle cannot become too flat as $h \to 0$ for a shape-regular family of triangulations.

We now state the main result.

Proposition 5.18 (polynomial interpolation error bound)**.** Let $\{\mathcal{T}_h\}_{h>0}$ be a family of shape-regular triangulations, and $\mathcal{I}_h w$ be the piecewise polynomial interpolant of degree p associated with $\mathcal{T}_h$. If $w \in C^0(\bar{\Omega}) \cap H^{s+1}(\mathcal{T}_h)$, then

$$\|w - \mathcal{I}_h w\|_{L^2(\Omega)} \leq C_{\mathcal{I}} h^{r+1} |w|_{H^{r+1}(\mathcal{T}_h)}$$
$$|w - \mathcal{I}_h w|_{H^1(\Omega)} \leq C'_{\mathcal{I}} h^{r} |w|_{H^{r+1}(\mathcal{T}_h)}$$

for $r = \min\{s, p\}$ and some $C_{\mathcal{I}}$ and $C'_{\mathcal{I}}$ independent of w and h.

Proof. Proof is beyond the scope of this course. We refer to Brenner and Scott (2008). □

The proposition summarizes the particular results we obtained in one dimension in Sections 5.4, 5.5, and 5.6. For functions in $C^0(\bar{\Omega}) \cap H^{p+1}(\mathcal{T}_h)$, the $L^2(\Omega)$ and $H^1(\Omega)$ norm of the errors associated with the piecewise degree-p interpolant converge as h^{p+1} and h^p, respectively. If the underlying function is not smooth but is only in $C^0(\bar{\Omega}) \cap H^{s+1}(\mathcal{T}_h)$ for $s < p$, then the convergence is limited to h^{s+1} in $L^2(\Omega)$ and h^s in $H^1(\Omega)$.

(The condition $C^0(\bar{\Omega}) \cap H^{s+1}(\mathcal{T}_h)$, instead of $H^1(\Omega) \cap H^{s+1}(\mathcal{T}_h)$, is necessary because $C^0(\bar{\Omega}) \subset H^1(\Omega)$ does not hold in $\mathbb{R}^{d>1}$; hence, the point-wise evaluation of a function, which is required to construct the interpolant, is in general ill-defined for $H^1(\Omega)$ functions in $\mathbb{R}^{d>1}$ (unlike in $\mathbb{R}^1$). This condition can be relaxed by considering so-called quasi-interpolants; however, the discussion is beyond the scope of this course.)

5.8 Isoparametric polynomial interpolation

To approximate domains with curved boundaries, we introduced in the previous lecture polynomial-based geometry mappings $\{\mathcal{G}^K\}_{K \in \mathcal{T}_h}$, where each $\mathcal{G}^K$ maps from a reference element $\tilde{K}$ to the physical (curved) element K. We then identified the associated approximation space by

$$\mathcal{V}_h \equiv \{v \in H^1(\Omega_h) \mid v|_K \circ \mathcal{G}^K \in \mathbb{P}^p(\tilde{K}),\ \forall K \in \mathcal{T}_h\}.$$

The curved elements yielding a better approximation for curved domains is perhaps rather intuitive. In fact, if a curved domain is approximated using non-curved elements, then we can readily show

that using higher-order ($\mathbb{P}^{p>1}$) finite elements is not asymptotically more accurate than using linear finite elements ($\mathbb{P}^{p=1}$). We now wish to understand the approximation properties of the interpolants associated with $\mathcal{V}_h$. To this end, we first extend the notion of shape-regular families of triangulations to curved meshes.

Definition 5.19 (shape-regular meshes (curved))**.** The family of curved meshes $\{\mathcal{T}_h\}_{h>0}$ is said to be shape-regular if it is shape-regular in the sense of Definition 5.17 and in addition the triangulation $\mathcal{T}_h \equiv \{K \equiv \mathcal{G}^K(\tilde{K})\}_{K\in\mathcal{T}_h}$ and the associated approximation of the domain $\Omega_h \equiv \cup_{K\in\mathcal{T}_h} K$ satisfy the following properties:

(i) The geometry mapping is affine for all elements not on the boundary; i.e., $\mathcal{G}^K \in \mathbb{P}^1(\tilde{K})$ for all K such that $\partial K \cap \partial\Omega_h = \emptyset$.

(ii) The distance from any point on $\partial\Omega$ to the closest point on $\partial\Omega_h$ is at most Ch^{p+1}.

(iii) $|J^K(\tilde{x})| \le C$ and $|J^K(\tilde{x})^{-1}| \le C$ for almost everywhere in $\tilde{K}$ for all $K \in \mathcal{T}_h$.

The second condition sets the minimum convergence rate at which the family of domains $\{\Omega_h\}_{h>0}$ must approximate the exact domain Ω as $h \to 0$. If the domain boundary is piecewise C^{p+1}, then the required rate can be achieved using $\mathbb{P}^p$ geometry mapping (assuming the triangulations match the kinks of the domain boundary). Third condition ensures that the geometry mapping does not become singular anywhere in the domain.

For a shape-regular triangulation of the curved domain, we obtain the following interpolation error bound.

Proposition 5.20 (isoparametric polynomial interpolation error bound)**.** Let $\Omega \subset \mathbb{R}^d$ be a curved domain, $\{\mathcal{T}_h\}_{h>0}$ be a family of shape-regular triangulations in the sense of Definition 5.19, and $\mathcal{I}_h w \in \mathcal{V}_h$ be an interpolant that belongs to

$$\mathcal{V}_h \equiv \{v \in H^1(\Omega_h) \mid v|_K \circ \mathcal{G}^K \in \mathbb{P}^p(\tilde{K}),\ \forall K \in \mathcal{T}_h\}.$$

If $w \in C^0(\bar{\Omega}) \cap H^{s+1}(\mathcal{T}_h)$, then

$$\|w - \mathcal{I}_h w\|_{L^2(\Omega)} \le C_{\mathcal{I}} h^{r+1} |w|_{H^{r+1}(\mathcal{T}_h)}$$
$$|w - \mathcal{I}_h w|_{H^1(\Omega)} \le C'_{\mathcal{I}} h^{r} |w|_{H^{r+1}(\mathcal{T}_h)}$$

for $r = \min\{s, p\}$ and some $C_{\mathcal{I}}$ and $C'_{\mathcal{I}}$ independent of w and h.

Proof. Proof is beyond the scope of this course. We refer to Brenner and Scott (2008). □

The proposition shows that we recover the optimal convergence rate on curved domains for shape-regular triangulations (based on isoparametric mapping) that rapidly approximate the domain shape. As noted above, if affine ($\mathbb{P}^1$) meshes are used to approximate curved boundaries, then the convergence rate for higher-degree interpolants is asymptotically the same as that for linear interpolants.

5.9 Summary

We summarize key points of this lecture:

1. The "classical" interpolation error bounds for C^k functions can be useful in many contexts but are not particularly well-suited for the analysis of finite element errors as it imposes a strong regularity requirement on the underlying function.

2. The Rayleigh quotient of a symmetric, positive bilinear form is bounded by the lower and upper bound of the associated eigenproblem.

3. For $w \in H^1(\Omega) \cap H^2(\mathcal{T}_h)$ and the associated piecewise linear interpolant $\mathcal{I}_h w$, $\|w - \mathcal{I}_h w\|_{L^2(\Omega)} \leq C_{\mathcal{I}} h^2 |w|_{H^2(\Omega)}$.

4. For $w \in H^1(\Omega) \cap H^2(\mathcal{T}_h)$ and the associated piecewise linear interpolant $\mathcal{I}_h w$, $|w - \mathcal{I}_h w|_{H^1(\Omega)} \leq C_{\mathcal{I}} h |w|_{H^2(\Omega)}$; the convergence rate is one lower than that for the $L^2(\Omega)$ error.

5. For $w \in H^1(\Omega)$ (but not $H^1(\Omega) \cap H^2(\mathcal{T}_h)$) and the associated piecewise linear interpolant $\mathcal{I}_h w$, $\|w - \mathcal{I}_h w\|_{L^2(\Omega)} \leq C_{\mathcal{I}} h |w|_{H^1(\Omega)}$; the convergence rate is one lower than that for a smoother function in $H^1(\Omega) \cap H^2(\mathcal{T}_h)$.

6. For $w \in C^0(\bar{\Omega}) \cap H^{s+1}(\Omega)$ and the associated piecewise degree-p polynomial interpolants on a family of shape-regular triangulations in $\mathbb{R}^d$, $\|w - \mathcal{I}_h w\|_{L^2(\Omega)} \leq C_{\mathcal{I}} h^{r+1} |w|_{H^{r+1}}$ and $|w - \mathcal{I}_h w|_{H^1(\Omega)} \leq C'_{\mathcal{I}} h^r |w|_{H^{r+1}}$ for $r = \min\{s, p\}$.

7. For a domain with a curved boundary, the above error bound holds assuming the family of shape-regular triangulations rapidly approximate the curved boundary using isoparametric mapping.

5.10 Appendix: Rayleigh quotient, Poincaré-Friedrichs inequality, and trace inequality

We recall the Poincaré-Friedrichs inequality, Proposition 2.35: for a Lipschitz domain $\Omega \subset \mathbb{R}^d$ with a boundary segment $\Gamma \subset \partial\Omega$ with $\Gamma \neq \emptyset$, there exists $C_{\rm PF} < \infty$ that only depends on Ω and Γ such that

$$\|v\|^2_{L^2(\Omega)} \leq C_{\rm PF}(|v|^2_{H^1(\Omega)} + \|v\|^2_{L^2(\Gamma)}) \quad \forall v \in H^1(\Omega).$$

Note that the constant $C_{\rm PF}$ can be expressed as

$$C_{\rm PF} = \sup_{v \in H^1(\Omega)} \frac{\|v\|^2_{L^2(\Omega)}}{|v|^2_{H^1(\Omega)} + \|v\|^2_{L^2(\Gamma)}}.$$

This is a Rayleigh quotient. The associated eigenproblem is as follows: find $(u_k, \lambda_k) \in H^1(\Omega) \times \mathbb{R}$ such that

$$(u_k, v)_{L^2(\Omega)} = \lambda_k (\int_\Omega \nabla v \cdot \nabla u_k dx + (u_k, v)_{L^2(\Gamma)}) \quad \forall v \in H^1(\Omega);$$

the Poincaré-Friedrichs constant is given by $C_{\rm PF} = \sup\{\lambda_k\}$.

We similarly recall the trace inequality, Proposition 2.39: for a Lipschitz domain $\Omega \subset \mathbb{R}^d$, there exists a constant $C_{\rm tr} < \infty$ that depends only on Ω such that

$$\|v\|_{L^2(\partial\Omega)} \leq C_{\rm tr} \|v\|_{H^1(\Omega)} \quad \forall v \in H^1(\Omega).$$

The square of the constant $C_{\rm tr}$ can be expressed as

$$C^2_{\rm tr} = \sup_{v \in H^1(\Omega)} \frac{\|v\|^2_{L^2(\partial\Omega)}}{\|v\|^2_{H^1(\Omega)}}.$$

This is again a Rayleigh quotient. The associated eigenproblem is as follows: find $(u_k, \lambda_k) \in H^1(\Omega) \times \mathbb{R}$ such that

$$(u_k, v)_{L^2(\partial\Omega)} = \lambda_k (u_k, v)_{H^1(\Omega)} \quad \forall v \in H^1(\Omega);$$

the constant $C_{\rm tr}$ is given by $C_{\rm tr} = \sup\{\lambda_k^{1/2}\}$.

Lecture 6

Finite element method: error analysis

6.1 Motivation

In this lecture, we analyze the error in finite element approximations. As we have seen in the previous lectures, the finite element method seeks a solution to the variational problem in (a family of) finite-dimensional approximation spaces, which often comprise piecewise polynomial functions. As such, the finite element error analysis builds on two distinct ingredients. The first is (quasi-)optimality results which show the ability of the Galerkin method to find a (quasi-)optimal approximation in a given finite-dimensional approximation space. The second is the approximation theory for the given approximation space; in the case of approximation spaces based on piecewise polynomials, we rely on the polynomial interpolation theory discussed in the previous lecture. The ability to carry out rigorous error analysis is one of the strengths of the finite element method, and we will demonstrate the strength in this lecture.

6.2 Preliminary

By way of preliminaries, we define equivalent norms.

Definition 6.1 (equivalence of norms)**.** Given a Hilbert space $\mathcal{V}$, a norm $\|\cdot\|_A$ is said to be equivalent to a norm $\|\cdot\|_B$ if there exist $c > 0$ and $C < \infty$ such that

$$c\|v\|_B \le \|v\|_A \le C\|v\|_B \quad \forall v \in \mathcal{V}.$$

We now introduce a set of assumptions used throughout this lecture. The first is a set of assumptions on the (abstract) variational problem.

Assumption 6.2. We consider the following.

1. The domain $\Omega \subset \mathbb{R}^d$ has a Lipschitz boundary.

2. The Hilbert space $\mathcal{V}$ satisfies $H_0^1(\Omega) \subset \mathcal{V} \subset H^1(\Omega)$. The space $\mathcal{V}$ is endowed with an inner product $(\cdot,\cdot)_{\mathcal{V}}$ and the associated induced norm $\|\cdot\|_{\mathcal{V}}$, which is equivalent to $\|\cdot\|_{H^1(\Omega)}$; i.e., $\exists C_{H^1\text{-}\mathcal{V}} < \infty$ and $C_{\mathcal{V}\text{-}H^1} < \infty$ such that

$$C_{H^1\text{-}\mathcal{V}}^{-1}\|v\|_{H^1(\Omega)} \le \|v\|_{\mathcal{V}} \le C_{\mathcal{V}\text{-}H^1}\|v\|_{H^1(\Omega)} \quad \forall v \in \mathcal{V}.$$

3. The bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is coercive and continuous in $\mathcal{V}$ with the coercivity and continuity constants $\alpha > 0$ and $\gamma < \infty$, respectively; i.e., $a(v,v) \ge \alpha\|v\|_{\mathcal{V}}^2$, $\forall v \in \mathcal{V}$, and $|a(w,v)| \le \gamma\|w\|_{\mathcal{V}}\|v\|_{\mathcal{V}}$ $\forall w, v \in \mathcal{V}$.

4. The linear form $\ell : \mathcal{V} \to \mathbb{R}$ is continuous in $\mathcal{V}$; i.e., $\exists c < \infty$ such that $|\ell(v)| \le c\|v\|_{\mathcal{V}}$ $\forall v \in \mathcal{V}$.

Assumption 6.2 does not assume the bilinear form is symmetric; we will clearly state the symmetry assumption whenever it is required as an additional assumption. We also note that Assumption 6.2 is a set of assumptions of the Lax-Milgram theorem, Theorem 2.32.

We next introduce the assumptions that define the variational solution and the associated finite element approximation.

Assumption 6.3. We consider the following.

1. The solution $u \in \mathcal{V}$ satisfies
$$a(u,v) = \ell(v) \quad \forall v \in \mathcal{V}. \tag{6.1}$$

2. The finite element approximation $u_h \in \mathcal{V}_h$ satisfies
$$a(u_h,v) = \ell(v) \quad \forall v \in \mathcal{V}_h \tag{6.2}$$
for some finite-dimensional subspace $\mathcal{V}_h \subset \mathcal{V}$.

Assumption 6.3 does not specify the finite element approximation space $\mathcal{V}_h$ other than that it is a subspace of $\mathcal{V}$; in particular, we do not assume the space $\mathcal{V}_h$ is a space of piecewise polynomials. Given Assumption 6.2, both the variational problem (6.1) and finite element problem (6.2) are well posed thanks to the Lax-Milgram theorem.

We finally introduce a particular family of piecewise polynomial approximation spaces.

Assumption 6.4. We consider the following:

1. The family of triangulations $\{\mathcal{T}_h\}$ is shape-regular in the in the sense of Definitions 5.17 and 5.19 for polygonal and curved domains, respectively.

2. The approximation spaces are given by
$$\mathcal{V}_h \equiv \{v \in \mathcal{V} \mid v \circ \mathcal{G}^K \in \mathbb{P}^p(\tilde{K}),\ K \in \mathcal{T}_h\}, \tag{6.3}$$
where $\{\mathcal{G}^K : \tilde{K} \to K\}_{K\in\mathcal{T}_h}$ is the geometry mapping associated with the shape-regular triangulations.

Note that (6.3) is one particular example of an approximation space for the finite element approximation (6.2). We will henceforth refer to the space $\mathcal{V}_h$ in (6.3) as the $\mathbb{P}^p$ finite element approximation space (even though the space may contain non-polynomial functions for isoparametric approximation of curved-domains). In addition, we will refer to the solution $u_h \in \mathcal{V}_h$ to (6.2) associated with $\mathcal{V}_h$ in (6.3) as the $\mathbb{P}^p$ finite element approximation.

6.3 Galerkin orthogonality

We now introduce *Galerkin orthogonality*, a relationship that will be used throughout our analysis of error in finite element approximations.

Lemma 6.5 (Galerkin orthogonality)**.** Suppose Assumptions 6.2 and 6.3 hold. The error $u-u_h \in \mathcal{V}$ satisfies

$$a(u - u_h, v) = 0 \quad \forall v \in \mathcal{V}_h.$$

Proof. The condition (6.1) implies $a(u, v) = \ell(v)$, $\forall v \in \mathcal{V}_h \subset \mathcal{V}$. The subtraction of (6.2) from the relationship yields

$$a(u - u_h, v_h) = a(u, v_h) - a(u_h, v_h) = \ell(v_h) - \ell(v_h) = 0 \quad \forall v_h \in \mathcal{V}_h,$$

which is the desired relationship. □

6.4 Error bounds in energy norm

In this section we consider a symmetric, coercive bilinear form and assess our error in *energy norm*.

Definition 6.6 (energy norm)**.** Given a symmetric, coercive, and continuous bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, the energy norm $\|\cdot\|_a : \mathcal{V} \to \mathbb{R}_{\geq 0}$ is defined by

$$\|v\|_a \equiv \sqrt{a(v, v)} \quad \forall v \in \mathcal{V}.$$

Because the bilinear form is symmetric and coercive, the bilinear form $a(\cdot, \cdot)$ is in fact an inner product that satisfies the requirements on (i) the linearity, (ii) symmetry, and (iii) the Cauchy-Shwarz inequality. The energy norm is the induced norm associated with this inner product; the norm hence satisfies the requirements on (i) linearity, (ii) positivity, and (iii) the triangle inequality. The energy norm is equivalent to $\|\cdot\|_{H^1(\Omega)}$:

Lemma 6.7 (equivalence of energy and H^1 norm)**.** Suppose Assumption 6.2 holds for a symmetric bilinear form. The energy norm $\|\cdot\|_a$ is equivalent to the $\mathcal{V}$ norm $\|\cdot\|_{\mathcal{V}}$, which in turn is equivalent to the H^1 norm $\|\cdot\|_{H^1(\Omega)}$.

Proof. From coercivity and continuity of $a(\cdot, \cdot)$ in $\mathcal{V}$, we immediately obtain

$$\alpha\|v\|_{\mathcal{V}}^2 \leq a(v, v) \equiv \|v\|_a^2 \leq \gamma\|v\|_{\mathcal{V}}^2 \quad \forall v \in \mathcal{V},$$

where $\alpha > 0$ and $\gamma < \infty$ are the coercivity and continuity constants, respectively. □

We now show that the finite element approximation is optimal in the energy norm.

Proposition 6.8 (energy-norm error bound)**.** Suppose Assumptions 6.2 and 6.3 hold for a symmetric bilinear form. The finite element approximation is optimal in the energy norm in the sense that

$$\|u - u_h\|_a = \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_a. \tag{6.4}$$

Proof. Let w_h be an arbitrary element in $\mathcal{V}_h$ and express it as $w_h = u_h + v_h$ for $v_h \in \mathcal{V}_h$. Then,

$$\begin{aligned}\|u - w_h\|_a^2 &= \|u - u_h - v_h\|_a^2 = a(u - u_h - v_h, u - u_h - v_h) \\ &= a(u - u_h, u - u_h) - 2 \underbrace{a(u - u_h, v_h)}_{=0 \text{ by Galerkin orthogonality}} + \underbrace{a(v_h, v_h)}_{>0 \text{ for } v_h \neq 0 \text{ by coercivity}} \\ &> \|u - u_h\|_a^2 \quad \forall v_h \neq 0,\end{aligned}$$

or, equivalently, $\|u - w_h\|_a > \|u - u_h\|_a \; \forall w_h \neq u_h$. □

The optimality of the finite element error in the energy norm implies the following: even *if* we knew the exact solution $u \in \mathcal{V}$ to (6.1), we could not find a $w_h \in \mathcal{V}_h$ that is more accurate in the energy norm than $u_h \in \mathcal{V}_h$. This optimality result is a direct consequence of Galerkin orthogonality, which states that the error $u - u_h \in \mathcal{V}$ is orthogonal to the space $\mathcal{V}_h$ in the inner product associated with the bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$. In other words, $u_h \in \mathcal{V}_h$ is the a-orthogonal projection of $u \in \mathcal{V}$ onto $\mathcal{V}_h \subset \mathcal{V}$.

We may obtain a particular h-convergence result for $\mathbb{P}^p$ finite element approximations.

Proposition 6.9 (energy-norm error bound: h convergence)**.** Suppose Assumptions 6.2, 6.3, and 6.4 hold for a symmetric bilinear form. If $u \in H^1(\Omega) \cap H^{s+1}(\mathcal{T}_h)$, then

$$\|u - u_h\|_a \leq C h^r |u|_{H^{r+1}(\mathcal{T}_h)}$$

for $r \equiv \min\{s, p\}$ and some constant $C < \infty$ independent of u and h. (Here, $H^k(\mathcal{T}_h)$ and $|\cdot|_{H^k(\mathcal{T}_h)}$ are the broken space and semi-norm, respectively, in Definition 5.10.)

Proof. The bound follows from the energy-norm error bound in Proposition 6.8 and the polynomial interpolation error bound in Proposition 5.18:

$$\begin{aligned}\|u - u_h\|_a &= \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_a && \text{(energy-norm error bound)} \\ &\leq \|u - \mathcal{I}_h u\|_a && (w_h = \mathcal{I}_h u) \\ &\leq \gamma \|u - \mathcal{I}_h u\|_{\mathcal{V}} && \text{(continuity of } a(\cdot,\cdot)) \\ &\leq \gamma C_{\mathcal{V}\text{-}H^1(\Omega)} \|u - \mathcal{I}_h u\|_{H^1(\Omega)} && \text{(equivalence of } \|\cdot\|_{\mathcal{V}} \text{ and } \|\cdot\|_{H^1(\Omega)}) \\ &\leq \gamma C_{\mathcal{V}\text{-}H^1(\Omega)} C_{\mathcal{I}} h^r |u|_{H^{r+1}(\mathcal{T}_h)}. && \text{(interpolation error bound)}\end{aligned}$$

We set $C \equiv \gamma C_{\mathcal{V}\text{-}H^1(\Omega)} C_{\mathcal{I}}$ to obtain the desired relationship. □

We observe that, if the solution u is smooth in the sense $u \in H^1(\Omega) \cap H^{p+1}(\mathcal{T}_h)$, then the error in the $\mathbb{P}^p$ finite element approximation converges as h^p in the energy norm. If the solution is not smooth in the sense $u \notin H^1(\Omega) \cap H^{p+1}(\mathcal{T}_h)$, then the convergence rate of the finite element approximation is limited by the regularity of the solution. We however note that the regularity of the solution is assessed in the broken norm $|u|_{H^{s+1}(\mathcal{T}_h)}$; if the irregular features in the solution, such as kinks, align with the triangulation, then the features may not deteriorate the convergence rate. Hence, for problems with known irregular features resulting from, say, discontinuous source functions or discontinuous diffusivity field, it is important to align the triangulation with the features.

6.5 Error bounds in $\mathcal{V}$ and $H^1(\Omega)$ norms

In this section, we obtain error bounds in the $\mathcal{V}$ and $H^1(\Omega)$ norms. As in Assumption 6.2, we assume $H^1_0(\Omega) \subset \mathcal{V} \subset H^1(\Omega)$ so that $\|\cdot\|_{\mathcal{V}}$ is equivalent to $\|\cdot\|_{H^1(\Omega)}$. The first bound is for a variational problem with a symmetric, coercive bilinear form.

Lemma 6.10 (Céa's lemma (symmetric))**.** Suppose Assumptions 6.2 and 6.3 hold for a symmetric bilinear form. Then,

$$\|u - u_h\|_{\mathcal{V}} \leq \sqrt{\frac{\gamma}{\alpha}} \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_{\mathcal{V}}. \tag{6.5}$$

Proof. The bound follows from the coercivity and continuity of the bilinear form and the energy-norm error bound in Proposition (6.8):

$$\begin{aligned}
\alpha\|u - u_h\|^2_{\mathcal{V}} &\leq a(u - u_h, u - u_h) && \text{(coercivity)} \\
&\leq \|u - u_h\|^2_a && \text{(energy norm)} \\
&= \inf_{w_h \in \mathcal{V}_h} \|u - u_h\|^2_a && \text{(energy-norm error bound)} \\
&= \gamma\|u - u_h\|^2_{\mathcal{V}} && \text{(continuity).}
\end{aligned}$$

The division by $\alpha > 0$ yields the desired inequality. □

The second bound is for a variational problem with a bilinear form that is coercive but not necessarily symmetric.

Lemma 6.11 (Céa's lemma (nonsymmetric))**.** Suppose Assumptions 6.2 and 6.3 hold. Then,

$$\|u - u_h\|_{\mathcal{V}} \leq \frac{\gamma}{\alpha} \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_{\mathcal{V}}. \tag{6.6}$$

Proof. The result is trivial for $\|u - u_h\|_{\mathcal{V}} = 0$. For $\|u - u_h\|_{\mathcal{V}} \neq 0$, we observe

$$\begin{aligned}
\alpha\|u - u_h\|^2_{\mathcal{V}} &\leq a(u - u_h, u - u_h) && \text{(coercivity)} \\
&= a(u - u_h, u - w_h) + a(u - u_h, w_h - u_h) && \text{(bilinearity)} \\
&= a(u - u_h, u - w_h) && \text{(Galerkin orthogonality)} \\
&\leq \gamma\|u - u_h\|_{\mathcal{V}}\|u - w_h\|_{\mathcal{V}} && \text{(continuity).}
\end{aligned}$$

The division by $\alpha\|u - u_h\|_{\mathcal{V}} > 0$ yields the desired result. □

Because $\gamma/\alpha \geq 1$ by the definition of the continuity and coercivity constants, the bound (6.6) for nonsymmetric bilinear forms, which applies to more general problems, is looser than the bound (6.5) for symmetric bilinear forms. In both cases, we observe that the finite element approximation is quasi-optimal in the sense that $\|u - u_h\|_{\mathcal{V}}$ is at most a constant multiple of the best-fit error $\inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_{\mathcal{V}}$, where the constant is independent of the approximation space $\mathcal{V}_h$.

Given the quasi-optimality results of Céa's lemma, we can readily obtain particular h-convergence results for $\mathbb{P}^p$ finite element approximations.

Proposition 6.12 ($\mathcal{V}$-norm error bound: h convergence)**.** Suppose Assumptions 6.2, 6.3, and 6.4 hold. If $u \in H^1(\Omega) \cap H^{s+1}(\mathcal{T}_h)$, then

$$\|u - u_h\|_{\mathcal{V}} \leq Ch^r |u|_{H^{r+1}(\mathcal{T}_h)} \tag{6.7}$$

for $r \equiv \min\{s, p\}$ and some $C < \infty$ independent of u and h.

Proof. We invoke Céa's lemma 6.10, the equivalence of $\|\cdot\|_{\mathcal{V}}$ and $\|\cdot\|_{H^1(\Omega)}$, and the polynomial interpolation error bound in Proposition 5.18. □

Proposition 6.13 ($H^1(\Omega)$-norm h convergence)**.** Suppose Assumptions 6.2, 6.3, and 6.4 hold. If $u \in H^1(\Omega) \cap H^{s+1}(\mathcal{T}_h)$, then

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch^r |u|_{H^{r+1}(\mathcal{T}_h)} \tag{6.8}$$

for $r \equiv \min\{s, p\}$ and some $C < \infty$ independent of u and h.

Proof. The result follows from Proposition 6.12 and the equivalence of $\|\cdot\|_{\mathcal{V}}$ and $\|\cdot\|_{H^1(\Omega)}$. □

Similar to the energy norm of the error, we observe that, if the solution u is smooth in the sense $u \in H^1(\Omega) \cap H^{p+1}(\mathcal{T}_h)$, then the error in $\mathbb{P}^p$ finite element approximations converges as h^p in the $\mathcal{V}$ or $H^1(\Omega)$ norm. If the solution is not smooth, then the convergence rate of the finite element approximations is limited by the regularity of the solution.

6.6 Error bounds in $L^2(\Omega)$ norm

We now analyze the convergence of finite element approximations in $L^2(\Omega)$ norm. Unfortunately, the $L^2(\Omega)$ error analysis relies on an equation-specific result called the *elliptic regularity estimate*. Hence, in this section, unlike in the previous sections, we restrict ourselves to (variable coefficients) advection-reaction-diffusion equation. (The regularity estimate holds also for other equations, but we here state a concrete result for the specific equation.)

Lemma 6.14 (elliptic regularity estimate)**.** Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, $\mathcal{V}$ be a Hilbert space such that $H_0^1(\Omega) \subset \mathcal{V} \subset H^1(\Omega)$, and let $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be

$$a(w, v) = \int_\Omega (\nabla v \cdot a \nabla w + v b \cdot \nabla w + cvw) dx, \quad \forall w, v \in \mathcal{V},$$

for $a \in C^1(\bar{\Omega})^{d \times d}$ and elliptic, $b \in C^0(\bar{\Omega})^d$, $c \in C^0(\bar{\Omega})$. Consider the following weak problem: find $u \in \mathcal{V}$ such that

$$a(u, v) = (f, v)_{L^2(\Omega)} \quad \forall v \in \mathcal{V},$$

where $f \in L^2(\Omega)$. Then the solution u satisfies

$$\|u\|_{H^2(\Omega)} \leq C_{\text{reg}} \|f\|_{L^2(\Omega)}$$

for some $C_{\text{reg}} < \infty$.

Proof. Proof is beyond the scope of this course. See, e.g., Ern and Guermond (2004). □

Proposition 6.15 (L^2-norm error bound (Aubin-Nitsche))**.** Suppose Assumptions 6.2, 6.3, and 6.4 as well as the conditions of the elliptic regularity estimate, Lemma 6.14, hold. Then,

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch\|u - u_h\|_{\mathcal{V}},$$

for some $C < \infty$ independent of u and h.

Proof. The proof is by so-called *Aubin-Nitsche trick*. We first pose a dual problem: find $\psi \in \mathcal{V}$ such that

$$a(w, \psi) = (w, e)_{L^2(\Omega)} \quad \forall w \in \mathcal{V}$$

for $e \equiv u - u_h$. We then observe that

$$\|e\|^2_{L^2(\Omega)} = a(e, \psi) = a(e, \psi - \mathcal{I}_h\psi) \leq \gamma\|e\|_{\mathcal{V}}\|\psi - \mathcal{I}_h\psi\|_{\mathcal{V}}.$$

We note that, since $u \in H^1(\Omega)$ and $u_h \in H^1(\Omega)$, $e \equiv u - u_h \in \mathcal{V} \subset H^1(\Omega) \subset L^2(\Omega)$; by the elliptic regularity estimate, $|\psi|_{H^2(\Omega)} \leq C_{\text{reg}}\|e\|_{L^2(\Omega)}$. We hence obtain

$$\begin{aligned}
\|\psi - \mathcal{I}_h\psi\|_{\mathcal{V}} &\leq C_{\mathcal{V}\text{-}H^1(\Omega)}\|\psi - \mathcal{I}_h\psi\|_{H^1(\Omega)} && \text{(equivalence of } \|\cdot\|_{\mathcal{V}} \text{ and } \|\cdot\|_{H^1(\Omega)}) \\
&\leq C_{\mathcal{V}\text{-}H^1(\Omega)}C_{\mathcal{I}}h|\psi|_{H^2(\mathcal{T}_h)} && \text{(interpolation error bound)} \\
&\leq C_{\mathcal{V}\text{-}H^1(\Omega)}C_{\mathcal{I}}C_{\text{reg}}h\|e\|_{L^2(\Omega)}. && \text{(elliptic regularity estimate)}
\end{aligned}$$

It follows

$$\|e\|^2_{L^2(\Omega)} \leq \gamma C_{\mathcal{V}\text{-}H^1(\Omega)}C_{\mathcal{I}}C_{\text{reg}}h\|e\|_{L^2(\Omega)}\|e\|_{\mathcal{V}}.$$

The division by $\|e\|_{L^2(\Omega)}$ yields the desired bound. □

Proposition 6.16 ($L^2(\Omega)$-norm error bound: h convergence)**.** Suppose Assumptions 6.2, 6.3, and 6.4 as well as the conditions of the elliptic regularity estimate, Lemma 6.14, hold. If $u \in H^1(\Omega) \cap H^{s+1}(\mathcal{T}_h)$, then

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{r+1}|u|_{H^{r+1}(\mathcal{T}_h)}$$

for $r \equiv \min\{s, p\}$ and some $C < \infty$ independent of u and h.

Proof. The result is a direct consequence of Propositions 6.15 and 6.12:

$$\begin{aligned}
\|u - u_h\|_{L^2(\Omega)} &\leq Ch\|u - u_h\|_{\mathcal{V}} && \text{(Aubin-Nitsche)} \\
&\leq C'h^{r+1}|u|_{H^{r+1}(\mathcal{T}_h)}. && (h \text{ convergence in } \|\cdot\|_{\mathcal{V}})
\end{aligned}$$

□

The proposition shows that, if the solution is smooth in the sense $u \in H^1(\Omega) \cap H^{p+1}(\mathcal{T}_h)$, then the $L^2(\Omega)$ norm of the error converges as h^{p+1} — the rate one higher than the $H^1(\Omega)$ or $\mathcal{V}$ norm of the error. In particular, we note that the $L^2(\Omega)$ norm of the error in the linear ($\mathbb{P}^1$) finite element approximations converge as h^2; because the $L^2(\Omega)$ norm is arguably the most popular metric for the assessment of the error in engineering, the linear finite element method is often quoted as a second-order method in the field.

6.7 Error bounds for functional outputs

In this section we consider the error in an *output* or *quantity of interest*. To begin, we introduce a linear functional associated with the output,

$$\ell^o : \mathcal{V} \to \mathbb{R};$$

we assume that the functional is continuous in $\mathcal{V}$: $\exists c < \infty$ such that $|\ell^o(w)| \leq c\|w\|_{\mathcal{V}}\ \forall w \in \mathcal{V}$.

In order to characterize output error, we first introduce the *dual problem*: find $\psi \in \mathcal{V}$ such that

$$a(w, \psi) = \ell^o(w) \quad \forall w \in \mathcal{V}. \tag{6.9}$$

The well-posedness of the dual problem follows from the Lax-Milgram theorem, Theorem 2.32, for a $\mathcal{V}$-coercive, $\mathcal{V}$-continuous (but not necessarily symmetric) bilinear form $a(\cdot,\cdot)$, and $\mathcal{V}$-continuous linear form $\ell^o(\cdot)$. The solution $\psi \in \mathcal{V}$ is called the *dual solution* or *adjoint*. (To contrast, the variational problem (6.1) is sometimes called the *primal problem* and the solution u is called the *primal solution*.)

Proposition 6.17 (output error bound (symmetric))**.** Suppose Assumptions (6.2) and (6.3) hold for a symmetric bilinear form. Let $\ell^o : \mathcal{V} \to \mathbb{R}$ be a continuous linear functional. Then,

$$|\ell^o(u) - \ell^o(u_h)| \leq \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_a \inf_{v_h \in \mathcal{V}_h} \|\psi - v_h\|_a,$$

where ψ is the solution to the dual problem 6.9.

Proof. We observe that, $\forall v_h \in \mathcal{V}_h$,

$$\begin{aligned}
|\ell^o(u) - \ell^o(u_h)| &= |\ell^o(u - u_h)| && \text{(linearity of } \ell^o\text{)} \\
&= |a(u - u_h, \psi)| && \text{(definition of adjoint } \psi\text{)} \\
&= |a(u - u_h, \psi - v_h)| && \text{(Galerkin orthogonality)} \\
&\leq \|u - u_h\|_a \|\psi - v_h\|_a && \text{(Cauchy-Schwarz)} \\
&\leq \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_a \|\psi - v_h\|_a. && \text{(energy-error optimality of } u_h\text{)}
\end{aligned}$$

We then take $v_h \in \mathcal{V}_h$ to be the minimizer of $\|\psi - v_h\|_a$ to obtain the desired result. □

Proposition 6.18 (output error bound (nonsymmetric))**.** Suppose Assumptions (6.2) and (6.3) hold. Let $\ell^o : \mathcal{V} \to \mathbb{R}$ be a continuous linear functional. Then,

$$|\ell^o(u) - \ell^o(u_h)| \leq \frac{\gamma^2}{\alpha} \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_{\mathcal{V}} \inf_{v_h \in \mathcal{V}_h} \|\psi - v_h\|_{\mathcal{V}},$$

where ψ is the solution to the dual problem 6.9.

Proof. We observe that, $\forall v_h \in \mathcal{V}_h$,

$$\begin{aligned}
|\ell^o(u) - \ell^o(u_h)| &= |\ell^o(u - u_h)| && \text{(linearity of } \ell^o\text{)} \\
&= |a(u - u_h, \psi)| && \text{(definition of adjoint } \psi\text{)} \\
&= |a(u - u_h, \psi - v_h)| && \text{(Galerkin orthogonality)} \\
&\leq \gamma \|u - u_h\|_{\mathcal{V}} \|\psi - v_h\|_{\mathcal{V}} && \text{(continuity of } a(\cdot,\cdot)\text{)} \\
&\leq \frac{\gamma^2}{\alpha} \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_{\mathcal{V}} \|\psi - v_h\|_{\mathcal{V}}. && (\|\cdot\|_{\mathcal{V}} \text{ error bound of } u_h)
\end{aligned}$$

We then take $v_h \in \mathcal{V}_h$ to be the minimizer of $\|\psi - v_h\|_{\mathcal{V}}$ to obtain the desired result. □

Proposition 6.19 (output error bound: h convergence)**.** Suppose Assumptions 6.2, 6.3, and 6.4 hold. Let $\ell^o : \mathcal{V} \to \mathbb{R}$ be a continuous linear functional. If $u \in H^1(\Omega) \cap H^{s+1}(\mathcal{T}_h)$ and $\psi \in H^1(\Omega) \cap H^{s'+1}(\mathcal{T}_h)$ for ψ the solution to the dual problem 6.9, then

$$|\ell^o(u) - \ell^o(u_h)| \leq Ch^{r+r'}|u|_{H^{r+1}(\mathcal{T}_h)}|\psi|_{H^{r'+1}(\mathcal{T}_h)}$$

for $r \equiv \min\{s, p\}$, $r' \equiv \min\{s', p\}$, and some constant $C < \infty$ independent of u, ψ, and h.

Proof. The result follows from (i) Proposition 6.18, (ii) the equivalence of $\|\cdot\|_{\mathcal{V}}$ and $\|\cdot\|_{H^1(\Omega)}$, and (iii) the polynomial interpolation error bound in Proposition 5.18. □

The proposition shows that for a smooth solution $u \in H^1(\Omega) \cap H^{p+1}(\mathcal{T}_h)$ *and* adjoint $\psi \in H^1(\Omega) \cap H^{p+1}(\mathcal{T}_h)$, the output converges as h^{2p}. The convergence rate for the output error is *twice* that for the $\mathcal{V}$ or $H^1(\Omega)$ norm of the error. This result is often referred to as *output superconvergence*. (The output superconverges because the finite element approximation is by construction *dual consistent*: the dual of the discrete problem is the discretization of the continuous dual problem. Not all discretizations for boundary value problems have this property.)

6.8 Generalization: other approximation spaces

Throughout this lecture, we presented two types of error bounds. First, under Assumptions 6.2 and 6.3, we obtained the (quasi-)optimality results that show the Galerkin finite element method achieves errors that are only some fixed constant away from the best-fit solution in a given approximation space (e.g., Céa's lemma, Lemma 6.11). Second, under Assumptions 6.2, 6.3, and 6.4, we obtained the particular h-convergence results for $\mathbb{P}^p$ finite element approximation spaces. The results of the first type, such as Céa's lemma

$$\|u - u_n\|_{\mathcal{V}} \leq \frac{\gamma}{\alpha} \inf_{w_n \in \mathcal{V}_n} \|u - w_n\|_{\mathcal{V}},$$

applies to any (family of) finite-dimensional approximation spaces $\{\mathcal{V}_n\}_{n>1}$. The Galerkin finite element method will find a quasi-optimal approximations $\{u_n\}_{n>1}$ in any family of approximation spaces $\{\mathcal{V}_n\}_{n>1}$.

We can derive various methods based on the Galerkin projection by choosing different approximation spaces. The "standard" finite element method based on h refinement considers $\{\mathcal{V}_h\}_{h>0}$ defined by Assumption 6.4; if the exact solution is in $H^1(\Omega) \cap H^{p+1}(\mathcal{T}_h)$, the $\mathcal{V}$-norm of the error converges as h^p. The spectral method considers approximation spaces consist of high-order global polynomials, $\mathcal{V}_p \equiv \{v \in \mathcal{V} \mid v \in \mathbb{P}^p(\Omega)\}$; if the exact solution is analytic, then the $\mathcal{V}$-norm of the error converges as $\exp(-Cp)$ for some C independent of p, achieving the so-called *exponential convergence*. The *hp* adaptive finite element method constructs a sequence of piecewise polynomial spaces of varying h and p tailored for the specific solution we wish to approximate. The extended finite element method (XEFM) or generalized finite element method (GFEM) considers a family of approximation spaces comprise specialized (non-polynomial) functions tailored for the specific features (e.g., corner singularity). The reduced-basis method, a model reduction method for parametrized PDEs, considers approximation spaces comprise specialized (non-polynomial) functions tailored for the parametric manifold. All of these techniques rely on the Galerkin projection, which identifies a quasi-optimal approximation in a given approximation space.

6.9 Summary

We summarize key points of this lecture:

1. For a symmetric, coercive problem, the energy norm is given by $\|\cdot\|_a \equiv \sqrt{a(\cdot,\cdot)}$. The finite element approximation is optimal in the energy norm in the sense that $\|u - u_h\|_a \leq \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_a$. If the solution is smooth, the energy norm of the error for the $\mathbb{P}^p$ finite element approximation converges as h^p.

2. Céa's lemma shows that the finite element approximation is quasi-optimal in the $\mathcal{V}$ norm in the sense that $\|u - u_h\|_{\mathcal{V}} \leq \frac{\gamma}{\alpha} \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_{\mathcal{V}}$. If the solution is smooth, the $\mathcal{V}$ norm of the error for the $\mathbb{P}^p$ finite element approximation converges as h^p.

3. The error bounds for the $H^1(\Omega)$ norm of the error is the same as that for the $\mathcal{V}$ norm of the error up to a constant.

4. If the solution is smooth, then the $L^2(\Omega)$ norm of the error for the $\mathbb{P}^p$ finite element approximation converges as h^{p+1}. The result follows from the Aubin-Nitsche trick.

5. The error in a linear functional output is a (scaled) product of the error in the primal and dual approximations. If both the primal and dual solutions are smooth, then the error in a linear functional output superconverges as h^{2p}.

6. For all of the above cases, if the solution is not smooth, then the converge rate may be limited by the regularity of the solution.

Lecture 7

Linear elasticity

7.1 Motivation

In this lecture we consider a weak formulation and the associated finite element approximation of linear-elasticity problems. Linear elasticity equations are of both historical significance and practical importance for finite element methods, as the methods were originally developed and are still used to address problems in structural mechanics. The linear elasticity equations also allow us to demonstrate the formulation and implementation of finite element methods for vector-valued equations.

7.2 Vector- and matrix-valued Sobolev spaces

In (steady-state) linear elasticity, we seek a vector-valued displacement field in $\Omega \subset \mathbb{R}^d$ that satisfies the Navier-Cauchy equations. By way of preliminaries, we introduce vector- and matrix-valued Sobolev spaces, which are required to describe the system of equations.

Definition 7.1 ($H^k(\Omega)^d$ space)**.** Given $\Omega \subset \mathbb{R}^d$ and an integer $k \geq 0$, a Hilbert space of vector-valued functions $H^k(\Omega)^d$ is endowed with an inner product

$$(w, v)_{H^k(\Omega)} \equiv \sum_{i=1}^{d} (w_i, v_i)_{H^k(\Omega)},$$

and the associated induced norm $\|w\|_{H^k(\Omega)} \equiv \sqrt{(w, w)_{H^k(\Omega)}}$; the space comprises functions

$$H^k(\Omega)^d \equiv \{v \mid \|v\|_{H^k(\Omega)} < \infty\}.$$

Here, v_i denotes the i-th component of the vector-valued field for $i = 1, \ldots, d$. In other words, for $v \in H^k(\Omega)^d$, we have $v : \Omega \to \mathbb{R}^d$ and $v_i \in H^k(\Omega)$, $i = 1, \ldots, d$. For $k = 0$, the space is denoted $L^2(\Omega)^d$. (Note that for notational brevity, we abbreviate $\|\cdot\|_{H^k(\Omega)^d}$ as $\|\cdot\|_{H^k(\Omega)}$.)

Definition 7.2 ($H^k(\Omega)^{d\times d}$ space)**.** Given $\Omega \subset \mathbb{R}^d$ and an integer $k \geq 0$, a Hilbert space of matrix-valued functions $H^k(\Omega)^{d\times d}$ is endowed with an inner product

$$(w, v)_{H^k(\Omega)} \equiv \sum_{i,j=1}^{d} (w_{ij}, v_{ij})_{H^k(\Omega)}$$

and the associated induced norm $\|w\|_{H^k(\Omega)} \equiv \sqrt{(w, w)_{H^k(\Omega)}}$; the space comprises functions

$$H^k(\Omega)^{d\times d} \equiv \{v \mid \|v\|_{H^k(\Omega)} < \infty\}.$$

Here, v_{ij} denotes the (i, j)-th component of the matrix-valued field for $i, j = 1, \ldots, d$. For $k = 0$, the space is denoted $L^2(\Omega)^{d\times d}$.

Definition 7.3 (dot product (vector field))**.** Given $w, v \in L^2(\Omega)^d$, the dot product $v \cdot w \in L^1(\Omega)$ such that

$$v \cdot w = \sum_{i=1}^{d} v_i w_i.$$

Definition 7.4 (dot product (matrix field))**.** Given $w, v \in L^2(\Omega)^d$, the dot product $v : w \in L^1(\Omega)$ such that

$$v : w = \sum_{i,j=1}^{d} v_{ij} w_{ij}.$$

Definition 7.5 (gradient of $H^1(\Omega)^d$ functions)**.** For $v \in H^1(\Omega)^d$, the gradient $\nabla v \in L^2(\Omega)^{d\times d}$ is a matrix-valued field such that

$$(\nabla v)_{ij} = \frac{\partial v_i}{\partial x_j}, \quad i, j = 1, \ldots, d.$$

Corollary 7.6. For $v \in H^1(\Omega)^2$, the gradient $\nabla v \in L^2(\Omega)^{2\times 2}$ is given by

$$\nabla v = \begin{pmatrix} \frac{\partial v_1}{\partial x_1} & \frac{\partial v_1}{\partial x_2} \\ \frac{\partial v_2}{\partial x_1} & \frac{\partial v_2}{\partial x_2} \end{pmatrix}.$$

Definition 7.7 (divergence of $H^1(\Omega)^d$ functions)**.** For $v \in H^1(\Omega)^d$, the divergence $\nabla \cdot v \in L^2(\Omega)$ is a scalar-valued field such that

$$\nabla \cdot v = \sum_{i=1}^{d} \frac{\partial v_i}{\partial x_i}.$$

Corollary 7.8. For $v \in H^1(\Omega)^2$, the divergence $\nabla \cdot v \in L^2(\Omega)$ is given by

$$\nabla \cdot v = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2}.$$

Definition 7.9 (divergence of $H^1(\Omega)^{d\times d}$ functions)**.** For $\sigma \in H^1(\Omega)^{d\times d}$, the divergence $\nabla \cdot \sigma \in L^2(\Omega)^d$ is a vector-valued field such that

$$(\nabla \cdot \sigma)_i = \sum_{i=1}^{d} \frac{\partial \sigma_{ij}}{\partial x_j}, \quad i = 1, \ldots, d.$$

Corollary 7.10. For $\sigma \in H^1(\Omega)^{2\times 2}$, the divergence $\nabla \cdot \sigma \in L^2(\Omega)^2$ is given by

$$\nabla \cdot \sigma = \begin{pmatrix} \frac{\partial \sigma_{11}}{\partial x_1} + \frac{\partial \sigma_{12}}{\partial x_2} \\ \frac{\partial \sigma_{21}}{\partial x_1} + \frac{\partial \sigma_{22}}{\partial x_2} \end{pmatrix}.$$

7.3 Variational formulation

We now formulate the linear elasticity problem. Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. We partition the boundary $\partial\Omega$ into a Dirichlet boundary Γ_D and a Neumann boundary Γ_N such that $\Gamma_D \cap \Gamma_N = \emptyset$ and $\partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_N$. We assume that the Dirichlet boundary is non-empty: $\Gamma_D \neq \emptyset$. Given a displacement field $v \in H^1(\Omega)^d$, we introduce the strain tensor (field) $\epsilon(v) \in L^2(\Omega)^{d\times d}$ such that

$$\epsilon(v) = \frac{1}{2}(\nabla v + \nabla v^T).$$

We next introduce the associated stress tensor (field). For an isotropic material, the stress field $\sigma(v) \in L^2(\Omega)^{d\times d}$ is given by

$$\sigma(v) = 2\mu\epsilon(v) + \lambda \mathrm{tr}(\epsilon(v))I,$$

where $\lambda \in L^\infty(\Omega)$ and $\mu \in L^\infty(\Omega)$ are the first and second Lamé parameters (fields), respectively, such that

$$\begin{aligned} 0 \leq \lambda(x) < \lambda_{\max} < \infty \quad \text{a.e. in } \Omega, \\ 0 < \mu_{\min} \leq \mu(x) \leq \mu_{\max} < \infty \quad \text{a.e. in } \Omega, \end{aligned}$$

and $\mathrm{tr}(A) \equiv \sum_{i=1}^d A_{ii}$ is the trace operator for any $A \in L^2(\Omega)^{d\times d}$. We now introduce the strong form of the linear elasticity problem: find u such that

$$\begin{aligned} -\nabla \cdot \sigma(u) &= f \quad \text{in } \Omega \\ u &= u^B \quad \text{on } \Gamma_D, \\ n \cdot \sigma(u) &= g \quad \text{on } \Gamma_N, \end{aligned}$$

where $f : \Omega \to \mathbb{R}^d$ is the body force field and $g : \Gamma_N \to \mathbb{R}^d$ is the traction force field. The first equation represents the force-equilibrium condition in the material. The second equation represents the prescribed displacement boundary condition. The third equation represents the traction (i.e., prescribed force) boundary condition.

We now derive a weak formulation of the linear elasticity problem. To this end, we first introduce a Hilbert space

$$\mathcal{V} \equiv \{v \in H^1(\Omega)^d \mid v|_{\Gamma_D} = 0\} \tag{7.1}$$

and an affine space

$$\mathcal{V}^E \equiv u^E + \mathcal{V},$$

where u^E is any function in $H^1(\Omega)^d$ such that $u^E|_{\Gamma_D} = u^B$. We recall that Dirichlet boundary conditions are essential boundary conditions that must be enforced explicitly through the choice of the space. While we here assume that the Dirichlet boundary condition is imposed on all d components on Γ_D for notational brevity, we can readily handle cases where a Dirichlet condition is imposed on some of the d components; this case arises, for instance, if a boundary is constrained from moving in the normal direction but can slide along the tangential directions.

We next take an arbitrary test function $v \in \mathcal{V}$, multiply the governing equation by v, integrate by parts, and make appropriate substitutions for the natural boundary conditions to obtain

$$\begin{aligned}
0 &= \int_\Omega v \cdot (-\nabla \cdot \sigma(u)) dx - \int_\Omega v \cdot f dx \\
&= \int_\Omega \nabla v : \sigma(u) dx - \underbrace{\int_{\Gamma_D} v \cdot (n \cdot \sigma(u)) ds}_{=0 \,:\, \text{Dirichlet BC}} - \int_{\Gamma_N} v \cdot \underbrace{(n \cdot \sigma(u))}_{g \,:\, \text{Neumann BC}} ds - \int_\Omega v \cdot f dx \\
&= \int_\Omega \nabla v : \sigma(u) dx - \int_{\Gamma_N} v \cdot g ds - \int_\Omega v \cdot f dx.
\end{aligned}$$

We can further simplify the term involving the integration over Ω. We first recall that $\sigma(u) = 2\mu\epsilon(u) + \lambda \text{tr}(\epsilon(u))I$. We next note that, because $\epsilon(u)$ is symmetric, $\nabla v : \epsilon(u) = \epsilon(v) : \epsilon(u)$. We then note that, because $\epsilon(\cdot)$ preserves the diagonal terms, $\nabla v : I = \text{tr}(\nabla v) = \text{tr}(\epsilon(v))$. It hence follows that

$$\nabla v : \sigma(u) = \nabla v : (2\mu\epsilon(u) + \lambda \text{tr}(\epsilon(u))I) = 2\mu\epsilon(v) : \epsilon(u) + \lambda \text{tr}(\epsilon(v))\text{tr}(\epsilon(u)).$$

Our weak formulation is as follows: find $u \in \mathcal{V}^E$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathcal{V}, \tag{7.2}$$

where

$$a(w, v) \equiv \int_\Omega (2\mu\epsilon(v) : \epsilon(w) + \lambda \text{tr}(\epsilon(v))\text{tr}(\epsilon(w))) dx \quad \forall w, v \in \mathcal{V}, \tag{7.3}$$

$$\ell(v) \equiv \int_\Omega v \cdot f dx + \int_{\Gamma_N} v \cdot g ds \quad \forall v \in \mathcal{V}; \tag{7.4}$$

we assume $f \in L^2(\Omega)^d$ and $g \in L^2(\Gamma_N)^d$. (These requirements can be relaxed to $f \in H^{-1}(\Omega)^d$ and $g \in H^{-1/2}(\Gamma_N)^d$.) We also note that the bilinear form is symmetric. In addition, noting that $\text{tr}(\epsilon(v)) = \nabla \cdot v$, we could obtain an alternative bilinear form:

$$a(w, v) = \int_\Omega (2\mu\epsilon(v) : \epsilon(w) + \lambda(\nabla \cdot v)(\nabla \cdot w)) dx \quad \forall w, v \in \mathcal{V}; \tag{7.5}$$

this form emphasizes that the divergence of the displacement field is penalized by the first Lamé parameter λ.

As we will see shortly, the bilinear form (7.3) (or (7.5)) is coercive and symmetric. Hence, we may also consider the minimization formulation. Let $J : \mathcal{V} \to \mathbb{R}$ such that

$$J(v) \equiv \frac{1}{2} a(v, v) - \ell(v) \quad \forall v \in \mathcal{V}. \tag{7.6}$$

Our minimization formulation is as follows: find $u \in \mathcal{V}$ such that

$$u = \arg\min_{w \in \mathcal{V}} J(w).$$

7.4 Well-posedness

We now wish to understand if a solution to the variational problem (7.2) exists and, if so, is unique. To this end, we verify the conditions of the Lax-Milgram theorem, and in particular the $\mathcal{V}$-coercivity of the bilinear form (7.3). (The continuity of the bilinear and linear forms are relatively straightforward to prove.)

The challenge in proving the coercivity of the bilinear form (7.3) lies in the fact that our strain operator $\epsilon : H^1(\Omega)^d \rightarrow L^2(\Omega)^{d\times d}$ has a non-trivial kernel. For instance, in $\mathbb{R}^2$, we can readily show that

$$\epsilon(v) = 0 \quad \forall v \in \mathcal{V}_{\text{RM}},$$

where

$$\mathcal{V}_{\text{RM}} \equiv \left\{ v \mid v = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + b \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix},\ a_1, a_2, b \in \mathbb{R} \right\}$$

is the space of infinitesimal rigid-body motion. In other words, we obtain zero strain for any displacement that is (i) rigid-body translation (described by a_1 and a_2) or (ii) (infinitesimal) rigid-body rotation (described by b). This is consistent with our physical interpretation of strain; rigid-body motion does not cause strain (or stress) in the material. This result for the linear elasticity equations can be contrasted to the result for Poisson's equation, where the kernel comprises only constant functions and the Poincaré-Friedrich's inequality was used to prove coercivity. The analysis of coercivity of the linear elasticity problem, which include also (infinitesimal) rigid-body rotation, requires the *Korn's inequality*.

Theorem 7.11 (Korn's inequality)**.** Let $\mathcal{V} \equiv \{v \in H^1(\Omega)^d \mid v|_{\Gamma_D} = 0\}$ with $\Gamma_D \neq \emptyset$. There exists $C_{\text{Korn}} > 0$ such that

$$\|\epsilon(v)\|_{L^2(\Omega)} \geq C_{\text{Korn}} \|v\|_{H^1(\Omega)} \quad \forall v \in \mathcal{V}.$$

Proof. Proof is beyond the scope of this course. We refer to Brenner and Scott (2008). □

We now show that the bilinear from (7.3) is coercive and continuous in $\mathcal{V}$, the linear form (7.4) is continuous in $\mathcal{V}$, and hence the solution to the weak problem (7.2) exists and is unique.

Proposition 7.12. The bilinear form (7.3) associated with the linear elasticity problem is symmetric, coercive, and continuous in $\mathcal{V}$ given by (7.1).

Proof. The symmetry of $a(\cdot,\cdot)$ is obvious from inspection. The coercivity of $a(\cdot,\cdot)$ is a consequence of the Korn's inequality: for any $v \in \mathcal{V}$

$$a(v,v) = 2\int_\Omega \mu\epsilon(v) : \epsilon(v)dx + \int_\Omega \lambda \text{tr}(\epsilon(v))^2 dx \geq 2\mu_{\min}\|\epsilon(v)\|_{L^2(\Omega)} \geq 2\mu_{\min} C_{\text{Korn}} \|v\|_{H^1(\Omega)}.$$

Hence $a(\cdot,\cdot)$ is coercive with the coercivity constant $\alpha \geq 2\mu_{\min} C_{\text{Korn}} > 0$. To show continuity we observe, $\forall w, v \in H^1(\Omega)^d$,

$$\begin{aligned}
|a(w,v)| &= 2\int_\Omega \mu\epsilon(v) : \epsilon(w)dx + \int_\Omega \lambda \text{tr}(\epsilon(v))\text{tr}(\epsilon(w))dx \\
&\leq 2\mu_{\max}\|\epsilon(v)\|_{L^2(\Omega)}\|\epsilon(w)\|_{L^2(\Omega)} + \lambda_{\max}\|\text{tr}(\epsilon(v))\|_{L^2(\Omega)}\|\text{tr}(\epsilon(w)\| \\
&\leq (2\mu_{\max} + \lambda_{\max})\|\epsilon(v)\|_{L^2(\Omega)}\|\epsilon(w)\|_{L^2(\Omega)} \\
&\leq (2\mu_{\max} + \lambda_{\max})\|v\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)};
\end{aligned}$$

here the last inequality follows from $\|\epsilon(v)\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)}$ because for any i and j,

$$\left(\frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)\right)^2 = \frac{1}{4}\left(\left(\frac{\partial v_i}{\partial x_j}\right)^2 + \left(\frac{\partial v_j}{\partial x_i}\right)^2 + 2\frac{\partial v_i}{\partial x_j}\frac{\partial v_j}{\partial x_i}\right) \leq \frac{1}{2}\left(\left(\frac{\partial v_i}{\partial x_j}\right)^2 + \left(\frac{\partial v_j}{\partial x_i}\right)^2\right),$$

where no sum is implied on the repeated indices. Hence $a(\cdot,\cdot)$ is continuous with the continuity constant $\gamma \leq 2\mu_{\max} + \lambda_{\max} < \infty$. □

Proposition 7.13. If $f \in L^2(\Omega)^d$ and $g \in L^2(\Gamma_N)^d$, then the linear form (7.4) associated with the linear elasticity problem is continuous in $\mathcal{V}$ given by (7.1).

Proof. We observe

$$\begin{aligned}
|\ell(v)| &= \left|\int_\Omega v \cdot f dx + \int_{\Gamma_N} v \cdot g ds\right| \\
&\leq \|v\|_{L^2(\Omega)}\|f\|_{L^2(\Omega)} + \|v\|_{L^2(\Gamma_N)}\|g\|_{L^2(\Gamma_N)} \\
&\leq \|v\|_{H^1(\Omega)}\|f\|_{L^2(\Omega)} + C_{\mathrm{tr}}\|v\|_{H^1(\Omega)}\|g\|_{L^2(\Gamma_N)} \\
&\leq (\|f\|_{L^2(\Omega)} + C_{\mathrm{tr}}\|g\|_{L^2(\Gamma_N)})\|v\|_{H^1(\Omega)}.
\end{aligned}$$

Hence $\ell(\cdot)$ is continuous with a continuity constant $c \leq \|f\|_{L^2(\Omega)} + C_{\mathrm{tr}}\|g\|_{L^2(\Gamma_N)} < \infty$. □

Proposition 7.14. The solution to the elasticity problem (7.2) exists and is unique.

Proof. By Propositions 7.12 and 7.13, the bilinear form (7.3) is coercive and continuous in $\mathcal{V}$, and the linear form (7.4) is continuous in $\mathcal{V}$. The existence and uniqueness of the solution follows from the Lax-Milgram theorem. □

7.5 Finite element method: formulation

To seek a finite element approximation, we introduced a vector-valued finite element space

$$\mathcal{V}_h \equiv \{v \in \mathcal{V} \mid v|_K \oplus \mathcal{G}^K \in \mathbb{P}^p(\tilde{K})^d,\ \forall K \in \mathcal{T}_h\},$$

where $\mathcal{G}^K : \tilde{K} \to K$ is the geometry mapping (for potentially curved domains) and $\mathbb{P}^p(K)^d$ is the space of vector-valued polynomials of degree p over K. We then consider the following finite element problem: find $u_h \in \mathcal{V}_h$ such that

$$a(u_h, v) = \ell(v) \quad \forall v \in \mathcal{V}_h \tag{7.7}$$

for the bilinear form (7.3) and the linear form (7.4). Because the bilinear form is coercive and continuous in $\mathcal{V}_h \subset \mathcal{V}$ and the linear form in continuous in $\mathcal{V}_h \subset \mathcal{V}$, the finite element problem has a unique solution by the Lax-Milgram theorem. In addition, we may consider the minimization formulation: find $u_h \in \mathcal{V}_h$ such that

$$u_h = \arg\min_{w_h \in \mathcal{V}_h} J(w_h),$$

where $J : \mathcal{V} \to \mathbb{R}$ is the functional defined in (7.6).

7.6 Finite element method: analysis

We can also readily analyze the error in the finite element approximation using the tools introduced in Lecture 6. Note in particular the linear elasticity problem (7.2) and the associated finite element problem (7.7) satisfy all the conditions of the Assumptions 6.2 and 6.3; in addition the bilinear form is symmetric.

To begin, we introduce the energy norm $\|\cdot\|_a \equiv \sqrt{a(\cdot,\cdot)}$; the energy norm of a displacement field for the linear elasticity problem is the total strain energy associated with the displacement field. We immediately obtain the optimality result in the energy norm: if $u \in \mathcal{V} \cap H^{s+1}(\mathcal{T}_h)^d$, then

$$\|u - u_h\|_a = \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_a \leq C h^r |u|_{H^{r+1}(\mathcal{T}_h)}$$

for $r \equiv \min\{s, p\}$ and some $C < \infty$ independent of u and h. (As discussed in Lecture 6, the result of the first type holds for any $\mathcal{V}_h \subset \mathcal{V}$, whereas the result of the second type is specific to the $\mathbb{P}^p$ finite element approximation space.) We also obtain a similar result in $H^1(\Omega)$ using the Céa's lemma: if $u \in \mathcal{V} \cap H^{s+1}(\mathcal{T}_h)^d$, then

$$\|u - u_h\|_{H^1(\Omega)} \leq \sqrt{\frac{\gamma}{\alpha}} \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_{H^1(\Omega)} \leq C h^r |u|_{H^{r+1}(\mathcal{T}_h)}$$

for $r \equiv \min\{s, p\}$ and some $C < \infty$ independent of u and h. It can also be shown that the elliptic regularity estimate holds for sufficiently regular domain and Lamé parameter fields, and hence the L^2 error can be analyzed using the Aubin-Nitsche trick: if $u \in \mathcal{V} \cap H^{s+1}(\mathcal{T}_h)^d$, then

$$\|u - u_h\|_{L^2(\Omega)} \leq C h^{r+1} |u|_{H^{r+1}(\mathcal{T}_h)}$$

for $r \equiv \min\{s, p\}$ and some $C < \infty$ independent of u and h. Finally, for a linear functional output $\ell^o(u)$, we obtain the output superconvergence result: if $u \in \mathcal{V} \cap H^{s+1}(\mathcal{T}_h)^d$ and $\psi \in \mathcal{V} \cap H^{s'+1}(\mathcal{T}_h)$, then

$$|\ell^o(u) - \ell^o(u_h)| \leq \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_a \inf_{v_h \in \mathcal{V}_h} \|\psi - v_h\|_a \leq C h^{r+r'} |u|_{H^{r+1}(\mathcal{T}_h)} |\psi|_{H^{r'+1}(\mathcal{T}_h)}$$

for $r = \min\{s, p\}$, $r' = \min\{s', p\}$, and some $C < \infty$ independent of u, ψ, and h. Here $\psi \in \mathcal{V}$ is the adjoint associated with the output functional $\ell^o(\cdot)$: $a(w, \psi) = \ell^o(w)\ \forall w \in \mathcal{V}$. One particular output that is often of interest is the *compliance output*, which results from $\ell^o \equiv \ell$ in (7.4). For the compliance output, $\psi = u$ and hence, if $u \in \mathcal{V} \cap H^{s+1}(\mathcal{T}_h)^d$,

$$|\ell(u) - \ell(u_h)| \leq \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_a^2 \leq C h^{2r} |u|^2_{H^{r+1}(\mathcal{T}_h)},$$

for $r = \min\{s, p\}$ and some $C < \infty$ independent of u and h.

7.7 Finite element method: implementation

We now discuss the implementation of finite element method. To this end, we introduce the space

$$H^1_h(\Omega) \equiv \{v \in H^1(\Omega) \mid v|_K \in \mathbb{P}^p(K), \quad \forall K \in \mathcal{T}_h\}$$

without any essential boundary conditions. We then introduce the associated Lagrange basis $\{\phi_k\}_{k=1}^m$. Note that the space (and the basis) are not vector-valued. Then, given a vector-valued function $v \in H_h^1(\Omega)^d$, we express its i-th component as

$$v_i(x) = \hat{v}_{ik}\phi_k(x) \quad \forall x \in \Omega,\ i = 1, \ldots, d,$$

for some $\hat{v} \in \mathbb{R}^{m\times d}$, with an implied sum on the repeated indices k. (The sum on repeated indices will be implied throughout this section unless stated otherwise.) Note that we have $d \cdot m$ coefficients because we must represent d different fields, each of which using m coefficients.

We now wish to identify the local stiffness matrix associated with an element $K \in \mathcal{T}_h$. To this end, we first rearrange the bilinear form (7.5) into a form more amenable to implementation:

$$\begin{aligned} a(w, v) &= \int_\Omega (\frac{1}{2}\mu(\nabla v + \nabla v^T) : (\nabla w + \nabla w^T) + \lambda(\nabla \cdot v)(\nabla \cdot w))dx \\ &= \int_\Omega (\mu \nabla v : \nabla w + \mu \nabla v : \nabla w^T + \lambda(\nabla \cdot v)(\nabla \cdot w))dx; \end{aligned}$$

here we have used the fact that $\nabla v^T : \nabla w = \nabla v : \nabla w^T$, $\nabla v^T : \nabla w^T = \nabla v : \nabla w$. We now evaluate the form for $v_i|_K = \hat{v}_{i\alpha}^K \phi_\alpha^K$ and $w_j|_K = \hat{w}_{j\beta}^K \phi_\beta^K$ to obtain

$$\begin{aligned} &a(w|_K, v|_K) \\ &= \int_K (\mu \frac{\partial v_i}{\partial x_j}\frac{\partial w_i}{\partial x_j} + \mu \frac{\partial v_i}{\partial x_j}\frac{\partial w_j}{\partial x_i} + \lambda \frac{\partial v_i}{\partial x_i}\frac{\partial v_j}{\partial x_j})dx \\ &= \int_K (\mu \hat{v}_{i\alpha}^K \frac{\partial \phi_\alpha^K}{\partial x_j}\frac{\partial \phi_\beta^K}{\partial x_j}\hat{w}_{i\beta}^K + \mu \hat{v}_{i\alpha}^K \frac{\partial \phi_\alpha^K}{\partial x_j}\frac{\partial \phi_\beta^K}{\partial x_i}\hat{w}_{j\beta}^K + \lambda \hat{v}_{i\alpha}^K \frac{\partial \phi_\alpha^K}{\partial x_i}\frac{\partial \phi_\beta^K}{\partial x_j}\hat{w}_{j\beta}^K)dx \\ &= \hat{v}_{i\alpha}^K \left(\int_K \mu \frac{\partial \phi_\alpha^K}{\partial x_j}\frac{\partial \phi_\beta^K}{\partial x_j}dx\right)\hat{w}_{i\beta}^K + \hat{v}_{i\alpha}^K \left(\int_K \mu \frac{\partial \phi_\alpha^K}{\partial x_j}\frac{\partial \phi_\beta^K}{\partial x_i}dx\right)\hat{w}_{j\beta}^K + \hat{v}_{i\alpha}^K \left(\int_K \lambda \frac{\partial \phi_\alpha^K}{\partial x_i}\frac{\partial \phi_\beta^K}{\partial x_j}dx\right)\hat{w}_{j\beta}^K. \end{aligned}$$

We recognize that the first term can be rearranged using a dummy index and Kronecker delta to obtain

$$a(w|_K, v|_K) = \hat{v}_{i\alpha}\hat{A}_{i\alpha j\beta}^K \hat{w}_{j\beta},$$

where the local stiffness matrix $\hat{A}^K \in \mathbb{R}^{(d\cdot n_s)\times(d\cdot n_s)}$ is given by

$$\hat{A}_{i\alpha j\beta}^K = \left(\sum_{k=1}^d \int_K \mu \frac{\partial \phi_\alpha}{\partial x_k}\frac{\partial \phi_\beta}{\partial x_k}dx\right)\delta_{ij} + \int_K \mu \frac{\partial \phi_\alpha}{\partial x_j}\frac{\partial \phi_\beta}{\partial x_i}dx + \int_K \lambda \frac{\partial \phi_\alpha}{\partial x_i}\frac{\partial \phi_\beta}{\partial x_j}dx.$$

It is convenient to think of the local stiffness matrix as a $d \times d$ block matrix; for instance, for $d = 2$,

$$\hat{A}^K = \left(\begin{array}{c|c} \hat{A}_{1,:,1,:}^K & \hat{A}_{1,:,2,:}^K \\ \hline \hat{A}_{2,:,1,:}^K & \hat{A}_{2,:,2,:}^K \end{array}\right),$$

where matrices $\hat{A}^K_{i,:,j,:} \in \mathbb{R}^{n_s \times n_s}$, $i, j = 1, 2$, are given by

$$\begin{aligned}
\hat{A}^K_{1,\alpha,1,\beta} &= \sum_{k=1}^{d} \int_K \mu \frac{\partial \phi_\alpha}{\partial x_k} \frac{\partial \phi_\beta}{\partial x_k} dx + \int_K \mu \frac{\partial \phi_\alpha}{\partial x_1} \frac{\partial \phi_\beta}{\partial x_1} dx + \int_K \lambda \frac{\partial \phi_\alpha}{\partial x_1} \frac{\partial \phi_\beta}{\partial x_1} dx, \\
\hat{A}^K_{1,\alpha,2,\beta} &= \int_K \mu \frac{\partial \phi_\alpha}{\partial x_2} \frac{\partial \phi_\beta}{\partial x_1} dx + \int_K \lambda \frac{\partial \phi_\alpha}{\partial x_1} \frac{\partial \phi_\beta}{\partial x_2} dx, \\
\hat{A}^K_{2,\alpha,1,\beta} &= \int_K \mu \frac{\partial \phi_\alpha}{\partial x_1} \frac{\partial \phi_\beta}{\partial x_2} dx + \int_K \lambda \frac{\partial \phi_\alpha}{\partial x_2} \frac{\partial \phi_\beta}{\partial x_1} dx, \\
\hat{A}^K_{2,\alpha,2,\beta} &= \sum_{k=1}^{d} \int_K \mu \frac{\partial \phi_\alpha}{\partial x_k} \frac{\partial \phi_\beta}{\partial x_k} dx + \int_K \mu \frac{\partial \phi_\alpha}{\partial x_2} \frac{\partial \phi_\beta}{\partial x_2} dx + \int_K \lambda \frac{\partial \phi_\alpha}{\partial x_2} \frac{\partial \phi_\beta}{\partial x_2} dx.
\end{aligned}$$

In this format, the pair of indices for the test function, $(i, \alpha) \in [1, d] \times [1, n_s]$, is mapped to a linear index $i \cdot n_s + \alpha \in [1, d \cdot n_s]$; similarly the pair of indices for the trial function, $(j, \beta) \in [1, d] \times [1, n_s]$, is mapped to a linear index $j \cdot n_s + \beta \in [1, d \cdot n_s]$.

Similarly, we can readily compute the local load vector. For $v_i|_K = \hat{v}^K_{i\alpha} \phi^K_\alpha$,

$$\ell(v|_K) = \int_K v_i f_i dx + \int_{\Gamma_N \cap \partial K} v_i g_i ds = \hat{v}_{i\alpha} \int_K \phi^K_\alpha f_i dx + \hat{v}_{i\alpha} \int_{\Gamma_N \cap \partial K} \phi^K_\alpha g_i ds.$$

We find that

$$\ell(v|_K) = \hat{v}_{i\alpha} \hat{f}^K_{i\alpha}$$

where the local load vector $\hat{f}^K \in \mathbb{R}^{(d \cdot n_s)}$ is given by

$$\hat{f}^K_{i\alpha} = \int_K \phi^K_\alpha f_i dx + \int_{\Gamma_N \cap \partial K} \phi^K_\alpha g_i ds.$$

It is again convenient to think of the local stiffness matrix as a d block vector; for instance, for $d = 2$,

$$\hat{f}^K = \begin{pmatrix} \hat{f}^K_{1,:} \\ \hline \hat{f}^K_{2,:} \end{pmatrix},$$

where $\hat{f}^K_{i,:} \in \mathbb{R}^{n_s}$, $i = 1, 2$, are given by

$$\begin{aligned}
\hat{f}^K_{1\alpha} &= \int_K \phi^K_\alpha f_1 dx + \int_{\Gamma_N \cap \partial K} \phi^K_\alpha g_1 ds, \\
\hat{f}^K_{2\alpha} &= \int_K \phi^K_\alpha f_2 dx + \int_{\Gamma_N \cap \partial K} \phi^K_\alpha g_2 ds.
\end{aligned}$$

A pair of indices for the test function, $(i, \alpha) \in [1, d] \times [1, n_s]$, is mapped to a linear index $i \cdot n_s + \alpha \in [1, d \cdot n_s]$. (In practice, the boundary term can be computed using the assembly technique for facet terms discussed in Section 4.6.3.)

We finally assemble the local stiffness matrices and load vectors to form a global stiffness matrix and load vector, respectively. To form the global stiffness matrix $\hat{A}_h \in \mathbb{R}^{(d \cdot m) \times (d \cdot m)}$, we successively insert the local stiffness matrices $\hat{A}^{K_k}_h \in \mathbb{R}^{(d \cdot n_s) \times (d \cdot n_s)}$ for $k = 1, \ldots, n_e$ according to

$$\hat{A}_{h,(ia)(jb)} \leftarrow \hat{A}_{h,(ia)(jb)} + \hat{A}^{K_k}_{(i\alpha)(j\beta)},$$

where $a = \theta_{K\text{-}n}(k, \alpha)$ and $b = \theta_{K\text{-}n}(k, \beta)$ for $\theta_{K\text{-}n}(\cdot, \cdot)$ the element-to-node mapping (i.e., connectivity). Similar to the local index, in practice the pairs of global indices $(i, a) \in [1, d] \times [1, m]$ and $(j, b) \in [1, d] \times [1, m]$ are mapped to linear global indices $i \cdot m + a$ and $j \cdot m + b$, respectively. To form the global load vector $\hat{f}_h \in \mathbb{R}^{(d \cdot m)}$, we successively insert the local load vectors $\hat{f}_h^{K_k} \in \mathbb{R}^{(d \cdot n_s)}$ for $k = 1, \ldots, n_e$ according to

$$\hat{f}_{h,(ia)} \leftarrow \hat{f}_{h,(ia)} + \hat{f}^{K_k}_{(i\alpha)},$$

where $a = \theta_{K\text{-}n}(k, \alpha)$; again, the global indices $(i, a) \in [1, d] \times [1, m]$ are mapped to linear global indices $i \cdot m + a$. We finally impose the essential (i.e. Dirichlet) boundary conditions following the procedure discussed in Section 4.7; we remove the rows and columns of $\hat{A}_h$ (and the columns of $\hat{f}_h$) associated with the degrees of freedom fixed by the essential boundary conditions. We then solve the linear system $\hat{A}_h \hat{u}_h = \hat{f}_h$ to obtain the coefficients $\hat{u}_h \in \mathbb{R}^{(d \cdot n)}$ associated with the displacement field (for non-Dirichlet nodes).

7.8 Nearly incompressible materials and locking for the $\mathbb{P}^1$ space

As analyzed in Section 7.6, the Galerkin finite element approximation provides a quasi-optimal approximation in $\mathcal{V}_h \subset \mathcal{V}$ for fixed Lamé parameters λ and μ. However, the performance of the $\mathbb{P}^1$ finite element method deteriorates as $\lambda \to \infty$; this phenomenon is known as *locking*. To observe the problem, we first recall the minimization problem: find $u \in \mathcal{V}$ such that

$$u = \arg\min_{w \in \mathcal{V}} \left(\frac{1}{2} \int_\Omega (2\mu \epsilon(w) : \epsilon(w) + \lambda (\nabla \cdot w)^2) dx - \int_\Omega w f dx - \int_{\Gamma_N} w g ds \right).$$

In the incompressible limit of $\lambda \to \infty$, the minimization problem becomes

$$u = \arg\min_{\substack{w \in \mathcal{V} \\ \nabla \cdot w = 0}} \left(\frac{1}{2} \int_\Omega 2\mu \epsilon(w) : \epsilon(w) dx - \int_\Omega w f dx - \int_{\Gamma_N} w g ds \right); \tag{7.8}$$

we observe that the solution must lie in the divergence-free space $\{v \in H^1(\Omega)^d \mid \nabla \cdot v = 0\}$ because the penalty on the divergence λ goes to ∞. However, for the $\mathbb{P}^1$ approximation space $\mathcal{V}_h = \{v \in H^1(\Omega)^d \mid v|_K \in \mathbb{P}^1(K)^d, \ \forall K \in \mathcal{T}_h \ ; \ v|_{\Gamma_D} = 0\}$ with the essential boundary condition, we can show

$$\{v \in \mathcal{V}_h \mid \nabla \cdot v = 0\} = \emptyset.$$

Because there is no nontrivial admissible member in the approximation space, the $\mathbb{P}^1$ finite element method does not converge in the incompressible limit. For λ finite but large (i.e., a nearly incompressible material), the quality of the $\mathbb{P}^1$ finite element approximation also deteriorates. Hence, for nearly incompressible materials, we must use either $\mathbb{P}^{p>1}$ finite elements or more exotic finite elements designed for divergence-free spaces.

(Note, the minimization equation for the incompressible limit, (7.8), is in fact also the equation governing the velocity field associated with incompressible Stokes flow, i.e., very viscous incompressible flow in the limit of vanishing inertia. The governing equation can be recast using a Lagrange multiplier — which is the pressure — as a saddle-point system. The solution to the saddle-point system can be approximated in appropriate finite element spaces; however, the approximation spaces for the velocity and pressure must be chosen to satisfy a stability condition known as the Brezzi-Babuška inf-sup condition.)

7.9 Summary

We summarize key points of this lecture:

1. Steady-state linear elasticity problems in $\mathbb{R}^d$ are governed by the Navier-Cauchy equations, which is a vector-valued equations with d components.

2. The weak formulation of the linear elasticity equations is cast in the space $\mathcal{V}$ such that $H_0^1(\Omega)^d \subset \mathcal{V} \subset H^1(\Omega)^d$ and yields a symmetric, coercive, and continuous bilinear form and a continuous linear form. In particular, the coercivity in $\mathcal{V}$ follows from the Korn's inequality assuming the Dirichlet boundary is nonempty to prevent rigid-body motions. The weak formulation is well-posed thanks to the Lax-Milgram theorem.

3. For any subspace $\mathcal{V}_h \subset \mathcal{V}$, the finite element approximation is well-posed thanks to the Lax-Milgram theorem.

4. The (quasi-)optimality of the Galerkin finite element approximations follows from the symmetry, coercivity, and continuity of the problem. For a smooth solution (and adjoint), the error converges as h^p in the energy norm and $H^1(\Omega)$ norm, h^{p+1} in the $L^2(\Omega)$ norm, and h^{2p} for outputs, including the compliance output. If the solution is not smooth, then the convergence rate may be limited by the regularity of the solution.

5. The implementation of a finite element solver for linear elasticity equations requires the extension of the implementation techniques developed in Lecture 4 to the system of equations.

Lecture 8

Adaptive finite element method

8.1 Motivation

We have so far considered finite element approximations where the triangulation and the associated approximation space are chosen *a priori* by the user. In this lecture we consider adaptive finite element methods, where a sequence of approximation spaces is constructed intelligently and automatically based on the solution behavior until a user-specified error tolerance is met (or the computational resource is exhausted). Adaptive finite element methods build on two key technical ingredients: the first is an *a posteriori* error estimation technique which allows the method to estimate, and also localize, the error in a given finite element approximation; the second is an adaptive mesh refinement strategy that refines appropriate elements based on the behavior of the (localized) error estimates.

8.2 Problem statement

Throughout this lecture we consider a general weak formulation of a PDE. To this end, we introduce a Lipschitz domain $\Omega \subset \mathbb{R}^d$, a Hilbert space $\mathcal{V}$ such that $H_0^1(\Omega) \subset \mathcal{V} \subset H^1(\Omega)$, an associated affine space $\mathcal{V}^E = u^E + \mathcal{V}$ where $u^E \in H^1(\Omega)$ satisfies the essential boundary conditions, a bilinear form $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$, a linear form $\ell : H^1(\Omega) \to \mathbb{R}$, and an output functional $\ell^o : H^1(\Omega) \to \mathbb{R}$. We assume that $a(\cdot,\cdot)$ is coercive and continuous in $\mathcal{V}$, and $\ell(\cdot)$ and $\ell^o(\cdot)$ are continuous in $\mathcal{V}$. We then consider the following weak statement: find $u \in \mathcal{V}^E$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathcal{V}, \tag{8.1}$$

then evaluate the output

$$s = \ell^o(\mu).$$

By the Lax-Milgram theorem, the solution to (8.1) exists and is unique. (We may also readily consider problems with a vector-valued solution field.)

We now consider a finite element approximation of (8.1). To this end, we introduce an approximation space

$$\mathcal{V}_h \equiv \{v \in \mathcal{V} \mid v|_K \in \mathbb{P}^p(K),\ \forall K \in \mathcal{T}_h\},$$

where $\mathcal{T}_h$ is a triangulation of Ω. (We may also readily consider isoparametric elements for curved domains.) Our finite element approximation is as follows: find $u_h \in \mathcal{V}_h$ such that

$$a(u_h, v) = \ell(v) \quad \forall v \in \mathcal{V}_h, \tag{8.2}$$

and evaluate the output

$$s_h = \ell^o(u_h).$$

The well-posedness of the problem follows from the coercivity and continuity of the bilinear form in $\mathcal{V}_h \subset \mathcal{V}$, the continuity of the linear form in $\mathcal{V}_h \subset \mathcal{V}$, and the Lax-Milgram theorem.

Our goal in *a posteriori* error estimation is to provide a computable estimate of the error in the field $\|u - u_h\|_{\mathcal{V}}$ (or output $|\ell^o(u) - \ell^o(u_h)|$ for some functional $\ell^o \in \mathcal{V}'$). Our goal in mesh adaptation is to identify a sequence of triangulations $\{\mathcal{T}_h\}_{h>0}$ that controls the error more efficiently than (say) uniform refinement.

8.3 Residual-based error estimate

8.3.1 Abstract formulation

By way of preliminaries, we consider an abstract form of error bounds and identify the key ingredients of an error bound. We first introduce the *residual* associated with a solution $u_h \in \mathcal{V}_h$: $r \in \mathcal{V}'$ such that

$$r(v) \equiv \ell(v) - a(u_h, v) \quad \forall v \in \mathcal{V}.$$

The residual is related to the error $e \equiv u - u_h$ by

$$a(e, v) = a(u, v) - a(u_h, v) = \ell(v) - a(u_h, v) = r(v) \quad \forall v \in \mathcal{V}.$$

We now appeal to (i) the coercivity of the bilinear form, (ii) the error-residual relationship, and (iii) the definition of the dual norm to obtain

$$\alpha \|e\|_{\mathcal{V}}^2 \leq a(e, e) = r(e) \leq \|r\|_{\mathcal{V}'} \|e\|_{\mathcal{V}},$$

or

$$\|e\|_{\mathcal{V}} \leq \frac{1}{\alpha} \|r\|_{\mathcal{V}'}. \tag{8.3}$$

We identify the two ingredients of our error bound: (1) the coercivity constant $\alpha = \inf_{v \in \mathcal{V}} \frac{a(v,v)}{\|v\|_{\mathcal{V}}^2}$; (2) the dual norm of the residual $\|r\|_{\mathcal{V}'} \equiv \sup_{v \in \mathcal{V}} \frac{r(v)}{\|v\|_{\mathcal{V}}}$. We discuss in the next sections computational approximation of these quantities.

8.3.2 Coercivity constant

We now consider the approximation of the first key ingredient of the error bound (8.3): the coercivity constant. We recall that the coercivity constant is given by

$$\alpha = \inf_{v \in \mathcal{V}} \frac{a(v, v)}{\|v\|_{\mathcal{V}}^2}, \tag{8.4}$$

which may be interpreted as the lower bound of the Rayleigh quotient associated with $a(\cdot,\cdot)$ and $(\cdot,\cdot)_{\mathcal{V}}$. The lower bound of the Rayleigh quotient is associated with the lower bound of the eigenvalues of the following eigenproblem: find $(u_k,\lambda_k)\in\mathcal{V}\times\mathbb{R}$, $k\in\mathbb{N}$, such that

$$\frac{1}{2}a(u_k,v)+\frac{1}{2}a(v,u_k)=\lambda_k(u_k,v)_{\mathcal{V}}\quad\forall v\in\mathcal{V}; \tag{8.5}$$

without loss of generality, we order the eigenvalues such that $\lambda_1\le\lambda_2\le\dots$ and identify the coercivity constant $\alpha=\inf_k\lambda_k$. (Here we do not assume $a(\cdot,\cdot)$ to be symmetric; the symmetrized eigenproblem (8.5) can be readily obtained as stationary points of the Lagrangian $\mathcal{L}(w,\mu)=a(w,w)-\mu((w,w)_{\mathcal{V}}-1)$ associated with (8.4).)

The solution to the coercivity eigenproblem (8.5) cannot be found in a closed form for a general $a(\cdot,\cdot)$. We can however consider the following finite-dimensional approximation of the coercivity constant in $\mathcal{V}_h\subset\mathcal{V}$:

$$\alpha_h\equiv\inf_{v\in\mathcal{V}_h}\frac{a(v,v)}{\|v\|_{\mathcal{V}}^2}. \tag{8.6}$$

The associated finite-dimensional eigenproblem is as follows: find $(u_{h,k},\lambda_{h,k})\in\mathcal{V}_h\times\mathbb{R}$, $k=1,\dots,n$, such that

$$\frac{1}{2}a(u_{h,k},v)+\frac{1}{2}a(v,u_{h,k})=\lambda_{h,k}(u_{h,k},v)_{\mathcal{V}}\quad\forall v\in\mathcal{V}_h;$$

again, without loss of generality, we order the eigenvalues such that $\lambda_{h,1}\le\cdots\le\lambda_{h,n}$ and identify the approximate coercivity constant $\alpha_h=\inf_k\lambda_{h,k}=\lambda_{h,1}$. The matrix form of the eigenproblem is as follows: find $(\hat{u}_{h,k},\lambda_{h,k})\in\mathbb{R}^n\times\mathbb{R}$, $k=1,\dots,n$, such that

$$\frac{1}{2}(\hat{A}_h+\hat{A}_h^T)\hat{u}_{h,k}=\lambda_{h,k}\hat{V}_h\hat{u}_{h,k}\quad\text{in }\mathbb{R}^n,$$

where $\hat{A}_h\in\mathbb{R}^{n\times n}$ is the stiffness matrix given by $\hat{A}_{h,ij}=a(\phi_j,\phi_i)$, and $\hat{V}_h\in\mathbb{R}^{n\times n}$ is the inner product matrix given by $\hat{V}_{h,ij}=(\phi_i,\phi_j)_{\mathcal{V}}$. The approximate coercivity constant α_h in (8.6) hence can be readily computed by solving the (finite-dimensional) eigenproblem using a (sparse) eigenproblem solver.

Unfortunately, the approximate coercivity constant α_h associated with $\mathcal{V}_h$ is an upper (and not lower) bound of the coercivity constant α associated with $\mathcal{V}$ because

$$\alpha_h\equiv\inf_{v_h\in\mathcal{V}_h}\frac{a(v_h,v_h)}{\|v_h\|_{\mathcal{V}}^2}\ge\inf_{v\in\mathcal{V}}\frac{a(v,v)}{\|v\|_{\mathcal{V}}^2}\equiv\alpha.$$

As a result, if we replace the coercivity constant α in the error bound (8.3) by α_h, then the resulting statement is no longer an error bound but merely an error estimate. However, under mild assumptions, it can also be shown that

$$|\lambda_1-\lambda_{h,1}|\le C\inf_{v_h\in\mathcal{V}_h}\|u_1-v_h\|_{\mathcal{V}}^2 \tag{8.7}$$

for some $C<\infty$. Hence, if the eigenproblem is approximated in a $\mathbb{P}^p$ space and the eigenfunction is in $\mathcal{V}\cap H^{p+1}(\mathcal{T}_h)$, then the eigenvalue superconverges as

$$|\lambda_1-\lambda_{h,1}|\le\tilde{C}h^{2p}|u_1|_{H^{p+1}(\Omega)}^2. \tag{8.8}$$

It follows that, even for a fairly coarse approximation space, we obtain a reasonable estimate of the minimum eigenvalue λ_1 and hence the coercivity constant α. In this lecture, we use α_h in place of α, accept the loss of the bound property for the simplicity of the implementation, and justify the choice by the convergence results in (8.7) and (8.8).

8.3.3 Dual norm of the residual: advection-reaction-diffusion equation

We now consider the approximation of the second key ingredient of the error bound (8.3): the dual norm of the residual. For simplicity, we consider the advection-reaction-diffusion equation associated with the space $\mathcal{V} \equiv \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}$ and

$$a(w,v) \equiv \int_\Omega (\nabla v \cdot \kappa \nabla w + vb \cdot \nabla w + cwv) dx \quad \forall w, v \in \mathcal{V},$$
$$\ell(v) \equiv \int_\Omega vf dx + \int_{\Gamma_N} vg ds \quad \forall v \in \mathcal{V}.$$

We first introduce the residual associated with elements and facets. The element residual $R_K \in L^2(K)$, $K \in \mathcal{T}_h$, is given by

$$R_K \equiv f + \nabla \cdot (\kappa \nabla u_h) - b \cdot \nabla u_h - c u_h; \tag{8.9}$$

note that R_K is the residual associated with the strong form of the equation. To define the facet residual, we first introduce a skelton of the triangulation $\mathcal{T}_h$, $\partial\mathcal{T}_h = \{F\}$, which comprises all facets of the triangulation. The facet residual $R_F \in L^2(F)$ is then given by

$$R_F \equiv \begin{cases} \frac{1}{2}[\![\kappa \nabla u_h]\!], & F \in \partial\mathcal{T}_h \setminus \partial\Omega, \\ n \cdot \kappa \nabla u_h - g, & F \in \Gamma_N, \\ 0, & F \in \Gamma_D, \end{cases} \tag{8.10}$$

where the jump operator on $F \in \partial\mathcal{T}_h \setminus \Gamma_N$ is given by

$$[\![\varphi]\!](x) = \lim_{\epsilon \to 0} (n^+ \cdot \varphi(x - \epsilon n^+) + n^- \cdot \varphi(x - \epsilon n^-)),$$

where n^+ and n^- are the outward-pointing unit normal vector from the two neighboring elements. (Note that the jump operator is independent of the particular assignment of the two elements to the "+" and "−" sides.) In short, the facet residual R_F is the jump in the diffusive flux for the internal facets and the misfit in the Neumann boundary condition for the facets on Γ_N.

We also introduce two constants that are required to estimate the dual norm of the residual $\|r\|_{\mathcal{V}'}$. The first constant is

$$\rho_K \equiv \sup_{v \in H^1(K)} \frac{\|v - \mathcal{I}_h v\|_{L^2(K)}}{\|v\|_{\mathcal{V}(K)}}, \tag{8.11}$$

where $\mathcal{I}_h v \in \mathcal{V}_h$ is an interpolant of $v \in \mathcal{V}$. The second constant is

$$\rho_{\partial K} \equiv \sup_{v \in H^1(K)} \frac{\|v - \mathcal{I}_h v\|_{L^2(\partial K)}}{\|v\|_{\mathcal{V}(K)}}. \tag{8.12}$$

The analytical solutions to these maximimization problem can be found only in limited cases. We can however estimate the constants by solving a finite-dimensional eigenproblems associated with the Rayleigh quotients (as done for the coercivity constant in Section 8.3.2).

We now bound the dual norm of the residual. To this end, we define $\mathcal{I}_h^r v \equiv v - \mathcal{I}_h v$ and note

$$
\begin{aligned}
&|r(v)| = |r(\mathcal{I}_h^r v)| \\
&= \left| \int_\Omega (\mathcal{I}_h^r v) f dx + \int_{\Gamma_N} (\mathcal{I}_h^r v) g ds - \int_\Omega (\nabla (\mathcal{I}_h^r v) \cdot \kappa \nabla u_h + (\mathcal{I}_h^r v) b \cdot \nabla u_h + (\mathcal{I}_h^r v) c u_h) dx \right| \\
&= \left| \sum_{K \in \mathcal{T}_h} \left(\int_K (\mathcal{I}_h^r v)(f + \nabla \cdot (\kappa \nabla u_h) - b \cdot \nabla u_h - c u_h) dx \right.\right. \\
&\qquad \left.\left. - \int_{\partial K \setminus \partial\Omega} (\mathcal{I}_h^r v) \frac{1}{2} [\![\kappa \nabla u_h]\!] ds - \int_{\partial K \cap \Gamma_N} (\mathcal{I}_h^r v)(n \cdot \kappa \nabla u_h - g) ds - \int_{\partial K \cap \Gamma_D} \cancelto{0}{(\mathcal{I}_h^r v)} (n \cdot \kappa \nabla u_h) ds \right) \right| \\
&= \left| \sum_{K \in \mathcal{T}_h} \left(\int_K (\mathcal{I}_h^r v) R_K dx - \int_{\partial K} (\mathcal{I}_h^r v) R_F ds \right) \right| \\
&\le \sum_{K \in \mathcal{T}_h} \left(\|\mathcal{I}_h^r v\|_{L^2(K)} \|R_K\|_{L^2(K)} + \|\mathcal{I}_h^r v\|_{L^2(\partial K)} \|R_F\|_{L^2(\partial K)} \right) \\
&\le \sum_{K \in \mathcal{T}_h} \underbrace{\left(\rho_K \|R_K\|_{L^2(K)} + \rho_{\partial K} \|R_F\|_{L^2(\partial K)} \right)}_{\eta_K} \|v\|_{\mathcal{V}(K)} \\
&\le \left(\sum_{K \in \mathcal{T}_h} \eta_K^2 \right)^{1/2} \left(\sum_{K \in \mathcal{T}_h} \|v\|_{\mathcal{V}(K)}^2 \right)^{1/2} \le \left(\sum_{K \in \mathcal{T}_h} \eta_K^2 \right)^{1/2} \|v\|_{\mathcal{V}}
\end{aligned}
$$

It hence follows that

$$
\|r\|_{\mathcal{V}'} \equiv \sup_{v \in \mathcal{V}} \frac{|r(v)|}{\|v\|_{\mathcal{V}}} \le \left(\sum_{K \in \mathcal{T}_h} \eta_K^2 \right)^{1/2} \equiv R_{\mathcal{T}_h},
$$

where

$$
\eta_K \equiv \rho_K \|R_K\|_{L^2(K)} + \rho_{\partial K} \|R_F\|_{L^2(\partial K)} \quad \forall K \in \mathcal{T}_h,
$$

where R_K, R_F, ρ_K, and $\rho_{\partial K}$ are defined by (8.9), (8.10), (8.11), and (8.12).

We make a few observations. First, if all quantities, and in particular ρ_K and $\rho_{\partial K}$, are computed exactly, then $R_{\mathcal{T}_h}$ bounds $\|r\|_{\mathcal{V}'}$ from the above. Second, in practice, because ρ_K and $\rho_{\partial K}$ are estimated, $R_{\mathcal{T}_h}$ is not a bound but merely an estimate. Third, the element-wise quantities $\{\eta_K\}_{K \in \mathcal{T}_h}$ can serve as elemental indicators with which we drive our mesh adaptation.

8.3.4 Output error estimate

In many engineering scenarios, our interest is in the prediction of a quantity of interest associated with a functional. We now wish to construct an error estimate and an associated local error indicator for the finite-element approximation of the output. To simplify the presentation, we consider a continuous linear functional $\ell^o \in \mathcal{V}'$ of the form

$$
\ell^o(w) \equiv \int_\Omega w f^o dx + \int_{\Gamma_N} w g^o ds \quad \forall w \in \mathcal{V};
$$

the first and second terms constitute volume and surface contributions to the output, respectively. We then introduce the adjoint problem: find $\psi \in \mathcal{V}$ such that

$$a(w, \psi) = \ell^o(w) \quad \forall w \in \mathcal{V}.$$

The associated finite element approximation is as follows: find $\psi_h \in \mathcal{V}_h$ such that

$$a(w, \psi_h) = \ell^o(w) \quad \forall w \in \mathcal{V}_h.$$

We also introduce the adjoint residual: $r^{\text{adj}} \in \mathcal{V}'$ such that

$$r^{\text{adj}}(w) \equiv \ell^o(w) - a(w, \psi_h) \quad \forall w \in \mathcal{V}.$$

We now appeal to (i) the definition of the adjoint, (ii) the Galerkin orthogonality, and (iii) the definition of the primal and adjoint residuals to obtain the following error bound: for $e \equiv u - u_h$ and $e^{\text{adj}} \equiv \psi - \psi_h$,

$$|\ell^o(u) - \ell^o(u_h)| = |a(e, \psi)| = |a(e, \psi - \psi_h)| = |r(e^{\text{adj}})| \leq \|r\|_{\mathcal{V}'} \|e^{\text{adj}}\|_{\mathcal{V}} \leq \frac{1}{\alpha} \|r\|_{\mathcal{V}'} \|r^{\text{adj}}\|_{\mathcal{V}'}.$$

This is our output error bound. However, the bound in general is not actionable because α, $\|r\|_{\mathcal{V}'}$ and $\|r^{\text{adj}}\|_{\mathcal{V}'}$ are not computable; we must estimate these quantities.

We have already discussed the estimation of the coercivity constant α by α_h in Section 8.3.2, and the estimation of the dual norm of the residual $\|r\|_{\mathcal{V}'}$ in Section 8.3.3. Using a technique similar to the one used in Section 8.3.3, we can also estimate the dual norm of the adjoint residual $\|r^{\text{adj}}\|_{\mathcal{V}'}$. To this end, we introduce the element adjoint residual

$$R_K^{\text{adj}} = f^o + \nabla \cdot (\kappa^T \nabla \psi_h) + \nabla \cdot (b \psi_h) - c \psi_h$$

and the facet adjoint residual

$$R_F^{\text{adj}} \equiv \begin{cases} \frac{1}{2} [\![\kappa^T \nabla \psi_h + b \psi_h]\!], & F \in \partial \mathcal{T}_h \setminus \partial \Omega \\ n \cdot (\kappa^T \nabla \psi_h + b \psi_h) - g^o, & F \in \Gamma_N, \\ 0, & F \in \Gamma_D. \end{cases}$$

Our estimate for $\|r\|_{\mathcal{V}'}$ is then given by

$$\|r^{\text{adj}}\|_{\mathcal{V}'} \leq \left(\sum_{K \in \mathcal{T}_h} (\eta_K^{\text{adj}})^2 \right)^{1/2} \equiv R_{\mathcal{T}_h}^{\text{adj}},$$

where

$$\eta_K^{\text{adj}} \equiv \rho_K \|R_K^{\text{adj}}\|_{L^2(K)} + \rho_{\partial K} \|R_F^{\text{adj}}\|_{L^2(\partial K)} \quad \forall K \in \mathcal{T}_h.$$

Finally, the local error indicator can be constructed by combining the primal and adjoint error indicator

$$\eta_K^o \equiv \eta_K \eta_K^{\text{adj}} \quad \forall K \in \mathcal{T}_h.$$

8.4 Extrapolation error estimate

8.4.1 Field error estimate

In this section we consider a simple but practical error estimate based on Richardson extrapolation. To illustrate the idea, we introduce two triangulations $\mathcal{T}_h$ and $\mathcal{T}_{h'}$, where $\mathcal{T}_{h'}$ results from a uniform refinement of $\mathcal{T}_h$ in which the linear length of each element is reduced by a factor of two; i.e., $h' = h/2$. We next introduce two $\mathbb{P}^p$ approximation spaces $\mathcal{V}_h$ and $\mathcal{V}_{h'}$ associated with $\mathcal{T}_h$ and $\mathcal{T}_{h'}$, respectively. We then introduce $\mathbb{P}^p$ finite element approximations, $u_h \in \mathcal{V}_h$ and $u_{h'} \in \mathcal{V}_{h'}$. We *assume* that the error in each solution varies as

$$\begin{aligned}\|u - u_h\|_{H^1(\Omega)} &\approx Ch^r \\ \|u - u_{h'}\|_{H^1(\Omega)} &\approx C(h')^r = Ch^r 2^{-r}\end{aligned}$$

for some constant $C < \infty$ and convergence rate r; we recall $r = p$ if $u \in H^{p+1}(\mathcal{T}_h)$, but $r < p$ for less regular solutions. To estimate the error in the *refined* solution $u_{h'} \in \mathcal{V}_{h'}$, we observe that

$$\begin{aligned}\|u_{h'} - u_h\|_{H^1(\Omega)} &= \|(u - u_h) - (u - u_{h'})\|_{H^1(\Omega)} \\ &\geq \|u - u_h\|_{H^1(\Omega)} - \|u - u_{h'}\|_{H^1(\Omega)} \qquad \text{(reverse triangle inequality)} \\ &\approx Ch^r(1 - 2^{-r}); \qquad \text{(h convergence estimate)}\end{aligned}$$

in the last step, we plausibly assume $\|u - u_h\|_{H^1(\Omega)} > \|u - u_{h'}\|_{H^1(\Omega)}$. It follows that

$$\|u - u_{h'}\|_{H^1(\Omega)} \approx Ch^r 2^{-r} = \frac{Ch^r 2^{-r}}{Ch^r(1 - 2^{-r})} Ch^r(1 - 2^{-r}) \lesssim \frac{1}{2^r - 1}\|u_{h'} - u_h\|_{H^1(\Omega)}.$$

Because $\|u_{h'} - u_h\|_{H^1(\Omega)}$ is computable, $\|u_{h'} - u_h\|_{H^1(\Omega)}/(2^r - 1)$ serves as a computable estimate of the error $\|u - u_{h'}\|_{H^1(\Omega)}$.

We can also localize the extrapolation error estimate to drive an adaptive finite element method. Our strategy is as follows: we adaptively refine the *coarse* mesh $\mathcal{T}_h$, but use the solution obtained on the *fine* solution $u_{h'} \in \mathcal{V}_{h'}$ as our current estimate of the solution (since presumably $u_{h'} \in \mathcal{V}_{h'}$ is more accurate than $u_h \in \mathcal{V}_h$). To this end, we define the local error indicator for an element $K \in \mathcal{T}_h$ in the coarse mesh as

$$\eta_K \equiv \|u_{h'} - u_h\|_{H^1(K)}.$$

In a practical implementation, it is convenient to first re-represent the solution $u_h \in \mathcal{V}_h$ in the refined space $\mathcal{V}_{h'}$ — we can represent the solution exactly because $\mathcal{V}_h \subset \mathcal{V}_{h'}$ — and then evaluate the integral in the refined space according to

$$\eta_K^2 = \sum_{K' \in \text{children}(K)} \|u_{h'} - u_h\|_{H^1(K')}^2,$$

where children$(K) \subset \mathcal{T}_{h'}$ is a set of child elements in $\mathcal{T}_{h'}$ that belongs to the parent element $K \in \mathcal{T}_h$. The local error indicator can then be used to mark elements with large errors to drive adaptive mesh refinement.

We make a few comments about our adaptive finite element method based on the extrapolation error estimate. First, it is a very simple procedure that builds directly on the two solutions rather than (say) the residual, the formulation is not equation specific and the same formulation can be

applied to any equations. Second, the procedure can be applied to any norms; for instance, the $L^2(\Omega)$ norm of the solution $u_{h'}$ may be approximated by $\|u - u_{h'}\|_{L^2(\Omega)} \lesssim \|u_{h'} - u_h\|_{L^2(\Omega)}/(2^r - 1)$, where $r = p + 1$ if $u \in H^{p+1}(\Omega)$. Third, while the procedure requires two solutions — one coarse and one fine —, the cost associated with the solution of the coarse problem is at least 2^d times smaller than the fine solution, where d is the dimensionality of the space, and in any event we use the fine solution as our approximation. Fourth, a reliable and effective error estimate requires an appropriate choice of the convergence rate r for the Richardson extrapolation; the overestimate of the convergence rate results in an underestimation of the error and vice versa. Fifth, a (very) conservative choice for r is to choose $r = 1/2$, which can provide a conservative but robust error estimate.

8.4.2 Output error estimate

The extrapolation error estimate can also be applied to assess the output error $|\ell^o(u) - \ell^o(u_{h'})|$ for some functional ℓ^o. By way of preliminaries, we recall that the output error depends on both the primal solution $u \in \mathcal{V}$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathcal{V},$$

and the dual solution $\psi \in \mathcal{V}$ such that

$$a(w, \psi) = \ell^o(w) \quad \forall w \in \mathcal{V}.$$

Specifically, if $u \in H^{s'+1}(\Omega)$ and $\psi \in H^{s''+1}(\Omega)$, then

$$|\ell^o(u) - \ell^o(u_h)| \leq \frac{\gamma^2}{\alpha} \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_{\mathcal{V}} \inf_{v_h \in \mathcal{V}_h} \|\psi - v_h\|_{\mathcal{V}} \leq C h^{r'+r''}$$

for $r' = \min\{s', p\}$ and $r'' = \min\{s'', p\}$. Hence, our extrapolation error estimate is given by

$$|\ell^o(u) - \ell^o(u_{h'})| \lesssim \frac{1}{2^r - 1} |\ell^o(u_{h'}) - \ell^o(u_h)|,$$

where $r = \min\{s', p\} + \min\{s'', p\}$.

We may also localize the output extrapolation error estimate to drive *output-based* or *goal-oriented* adaptive mesh refinement. We note that the output error depends on both the primal and dual solutions and define an output error indicator as

$$\eta_K \equiv \|u_{h'} - u_h\|_{H^1(K)} \|\psi_{h'} - \psi_h\|_{H^1(K)}.$$

Here, $\psi_h \in \mathcal{V}_h$ is the finite element approximation of the dual problem and satisfies

$$a(w, \psi_h) = \ell^o(w) \quad \forall w \in \mathcal{V}_h;$$

the approximation $\psi_{h'} \in \mathcal{V}_{h'}$ is defined in an analogous manner. It is important to note that this output error indicator marks elements that have a large primal *and* dual solution errors; if either error is small, even if the other one is large, then the element is not marked for refinement.

8.5 Adaptive mesh refinement

8.5.1 General procedure

Once we have a means to localize the error in the solution, we can drive mesh adaptivity. The standard "loop" employed in an adaptive finite element method is

$$\textsc{Solve} \;\rightarrow\; \textsc{Estimate} \;\rightarrow\; \textsc{Mark} \;\rightarrow \textsc{Refine},$$

which is repeated until the desired error tolerance is met. We first provide a brief description of each step for a general error estimation and error localization technique:

- Solve. This step solves the finite element problem on a given mesh to compute the finite element solution u_h.
- Estimate. This step estimates the solution error (say) $\|u - u_h\|_{\mathcal{V}}$ using an *a posteriori* error estimate. If the desired error tolerance is met, then the adaptation iteration is terminated.
- Mark. This step marks elements to be refined based on the local error estimates $\{\eta_K\}_{K\in\mathcal{T}_h}$. There are many different marking strategies; arguably the simplest strategy is the fixed-fraction marking strategy, which marks a given fixed fraction (say 10%) of the elements with the largest η_K for refinement. Elements can also be marked for coarsening, if the mesh structure supports coarsening.
- Refine. This step refines the marked elements and modifies the mesh. There are many different refinement strategies; arguably the simplest strategy is to isotropically subdivide the element. For instance, a line is split into two lines, and a triangle is split into four triangles. Additional refinements of unmarked neighbor elements may need to be performed in $\mathbb{R}^{d>1}$ if we wish to maintain a conforming mesh. (Some adaptive solvers support *hanging nodes*, where two neighboring elements can have different levels of refinement, and a node can "hang" in the middle of the shared facet.)

The steps are repeated until the desired error tolerance is met. The adaptive procedure yields a sequence of triangulations $\{\mathcal{T}_h\}$ and the associated approximation spaces $\{\mathcal{V}_h\}$ that are tailored for the particular solution u.

8.5.2 Adaptation for extrapolation error estimate

We now outline the adaptation procedure to estimate and control the $H^1(\Omega)$ norm of the error using the extrapolation error estimate. Throughout the description, $\mathcal{V}_h$ and $\mathcal{V}_{h'}$ are the $\mathbb{P}^p$ finite element approximation spaces associated with $\mathcal{T}_h$ and $\mathcal{T}_{h'}$, respectively.

0. Prepare an initial coarse mesh $\mathcal{T}_h^{(1)}$. Set the iteration counter to $i = 1$.
1. Uniformly refine the mesh $\mathcal{T}_h^{(i)}$ to obtain $\mathcal{T}_{h'}^{(i)}$.
2. Solve for the finite element solutions $u_h^{(i)} \in \mathcal{V}_h^{(i)}$ and $u_{h'}^{(i)} \in \mathcal{V}_{h'}^{(i)}$.
3. Re-represent the solution $u_h \in \mathcal{V}_h^{(i)}$ in the refined space $\mathcal{V}_{h'}^{(i)}$ such that error estimate and local error indicators can be computed as algebraic operations in the fine space.

4. ESTIMATE the error in the refined solution as $\|u - u_{h'}^{(i)}\|_{H^1(\Omega)} \lesssim \|u_{h'}^{(i)} - u_h^{(i)}\|_{H^1(\Omega)}/(2^r - 1)$. If the target error tolerance is met, terminate.

5. MARK the top $\alpha\%$ of elements with the largest error indicator $\eta_K \equiv \|u_{h'} - u_h\|_{H^1(K)}$ for refinement.

6. REFINE marked elements.

7. Set $i \leftarrow i + 1$, and go to Step 1.

The procedure is repeated until the target error tolerance is met.

We make two remarks. First, the above procedure can be readily adopted to estimate and control the error in a functional output by incorporating the output extrapolation error estimate described in Section 8.4.2. Second, this simple procedure can yield a sequence of adapted meshes that can significantly improve the efficiency of the finite element method in the presence of singularities and singular perturbations as outlined in Sections 8.6.2 and 8.6.3.

8.6 Adaptive mesh refinement and singularity

8.6.1 Regularity of Poisson solutions in $\mathbb{R}^2$

To illustrate the importance of mesh adaptation, we first analyze the regularity of the solution to a Poisson problem on a two-dimensional, pie-shaped domain with a corner

$$\Omega \equiv \{x = (r\cos(\theta), r\sin(\theta)) \in \mathbb{R}^2 \mid 0 \leq r < 1 \text{ and } 0 < \theta < \omega\};$$

note that the corner angle is $\omega \in (0, \pi)$. We then consider a Poisson problem:

$$\begin{aligned} -\Delta u &= 0 \quad \text{in } \Omega, \\ u(r, \theta = 0) &= u(r, \theta = \omega) = 0, \\ u(r = 1, \theta) &= \sin(\frac{\pi\theta}{\omega}). \end{aligned}$$

The solution to this Poisson problem is given by

$$u(r, \theta) = r^{\frac{\pi}{\omega}} \sin(\frac{\pi\theta}{\omega}).$$

We can readily verify that this solution satisfies the PDE: we recall that

$$-\Delta u = -\frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial u}{\partial r}) - \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}$$

and note that

$$\begin{aligned} \frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial u}{\partial r}) &= \left(\frac{\pi}{\omega}\right)^2 r^{\frac{\pi}{\omega}-2} \sin(\frac{\pi\theta}{\omega}), \\ \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} &= -\left(\frac{\pi}{\omega}\right)^2 r^{\frac{\pi}{\omega}-2} \sin(\frac{\pi\theta}{\omega}), \end{aligned}$$

which cancel each other to yield $-\Delta u = 0$. We also readily verify that the boundary conditions are satisfied: $u(r, \theta = 0) = u(r, \theta = \omega) = 0$ and $u(r = 1, \theta) = \sin(\pi\theta/\omega)$. We now assess the regularity of the solution for domains with various angles.

- *Quadrant ($\omega = \pi/2$).* We first consider the quadrant. The solution is given by $u(r,\theta) = r^2 \sin(2\theta)$. We readily observe that the $L^2(\Omega)$ norm of the solution is finite. Through tedious but straightforward manipulation, we can also show that the $H^1(\Omega)$ semi-norm of the solution is given by

$$|u|^2_{H^1(\Omega)} \equiv \int_\Omega \nabla u \cdot \nabla u dx = \int_{\theta=0}^{\omega} \int_{r=0}^{1} \left(\left(\frac{\partial u}{\partial r} \right)^2 + \left(\frac{1}{r} \frac{\partial u}{\partial \theta} \right)^2 \right) r dr d\theta = \frac{\pi}{2}.$$

  Similarly, the $H^2(\Omega)$ semi-norm of the solution is given by $|u|^2_{H^2(\Omega)} = \pi$. It hence follows that the solution is (at least) in $H^2(\Omega)$.

- *Crack ($\omega = 2\pi$).* We next consider a crack. The solution is given by $u(r,\theta) = r^{1/2} \sin(\theta/2)$. We again readily confirm that both the $L^2(\Omega)$ norm and the $H^1(\Omega)$ semi-norm of the solution are finite: in particular, $|u|_{H^1(\Omega)} = \pi/2$; hence $u \in H^1(\Omega)$. However, in this case, the $H^2(\Omega)$ norm of the solution is not finite; hence $u \notin H^2(\Omega)$.

- *General.* In general, it can be shown that $u \in H^{s+1}(\Omega)$ if and only if $s \in (0, \pi/\omega)$.

We now assess the impact of the regularity on the convergence rate. As we have seen before, for the piecewise $\mathbb{P}^p$ polynomial approximation space V_h associated with a uniform mesh of element diameter h, the best-fit approximation error is bounded by

$$\inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_{H^1(\Omega)} \leq C h^r |u|_{H^{1+r}(\Omega)},$$

where $r = \min\{p, s\}$. To provide a concrete example, consider the L-shaped domain: $\omega = 3\pi/2$. In this case, we have $s < 2/3$ and the convergence rate for the $H^1(\Omega)$ error is $2/3$ for any polynomial approximation ($p \geq 1$). We observe that the convergence rate is limited by the regularity of the solution, and not the choice of the polynomial. We may hence conclude that the use of a higher-degree polynomials is not beneficial. This is a true statement for *uniform meshes*; however, as we will see shortly, the $\mathbb{P}^{p>1}$ approximation can be effectively if we consider adaptive mesh refinement.

8.6.2 Singularity in $\mathbb{R}^1$

Adaptive mesh refinement in general can improve the convergence of the error (say) $\|u - u_h\|_{H^1(\Omega)}$ with respect to the number of degrees of freedom n. (Note that, instead of the maximum element diameter $h \equiv \max_{K \in \mathcal{T}_h} h_K$, we use the number of degrees of freedom $n \equiv \dim(\mathcal{V}_h)$ as the indicator of complexity, which is more suitable for non-uniform meshes.) To demonstrate the benefit, in this section, we consider a canonical singular solution in $\mathbb{R}^1$,

$$u(x;\alpha) = x^\alpha, \quad x \in \Omega \equiv (0,1),$$

for some parameter $\alpha > 1/2$. We recall from Section 8.6.1 this x^α-type singularity is encountered at corners of the domain in $\mathbb{R}^{d=2}$. We can readily show that $u(\cdot;\alpha)$ is in $H^s(\Omega)$ for $\alpha > s - 1/2$, but not for $\alpha = s - 1/2$ for $\alpha \notin \mathbb{Z}$.

We first consider the approximation of the singular solution in $\mathbb{P}^p$ approximation space associated with uniform meshes for $p \geq \alpha - 1/2$. (In practice, α is often in $(1/2, 1)$, and hence the condition $p \geq \alpha - 1/2$ is satisfied even for $p = 1$.) We can show that the error converges as

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^{\alpha - 1/2} = C n^{-\alpha + 1/2}. \tag{8.13}$$

The convergence rate is limited the regularity of the solution. In particular, if $u(x) = x^{1/2+\epsilon}$ for ϵ small, then $u \in H^1(\Omega)$ and the convergence rate is $n^{-\epsilon}$; we may observe an arbitrary slow convergence regardless of the choice of the approximation degree p.

We now consider a graded mesh whose element diameter varies in $\Omega \equiv (0,1)$ according to

$$h(x) \approx cx^{\beta}$$

for some $c > 0$ and a grading parameter $\beta > 0$. Note that the elements become smaller towards the singularity at $x = 0$. If the grading parameter β is optimally chosen for a given singularity strength α and polynomial degree p, then we can show that

$$\|u - u_h\|_{H^1(\Omega)} \leq C' n^{-p}. \tag{8.14}$$

In words, we recover the optimal convergence rate *with respect to the number of degrees of freedom* n observed for smooth solutions.

The comparison of the convergence results for the uniform mesh (8.13) and the graded mesh (8.14) highlights the importance of mesh adaptation for problems with singularities. This is particularly the case of for higher-order ($p > 1$) approximations, where the efficiency of higher-order method is realized only on appropriately graded meshes in the presence of singularities.

8.6.3 Singular perturbation in $\mathbb{R}^1$

The next example we consider is a canonical singular perturbation solution in $\mathbb{R}^1$,

$$u(x;\epsilon) = \exp(-x/\epsilon), \quad x \in (0,1),$$

for some ϵ such that $0 < \epsilon \ll 1$. This is the boundary-layer solution encountered in both reaction-diffusion and advection-diffusion equation with a weak diffusion (i.e., high Peclet/Reynolds number flows). Unlike the x^{α}-type singularity, this solution is formally smooth (i.e., infinitely differentiable). However, the solution exhibit rapid variation in the thin layer of $\mathcal{O}(\epsilon)$ in the vicinity of $x = 0$.

We first consider the approximation of the boundary layer solution in $\mathbb{P}^p$ approximation spaces associated with uniform meshes. Because the solution is formally smooth, the error asymptotically behaves as

$$\|u - u_h\|_{H^1(\Omega)} \leq C_\epsilon h^p \leq C_\epsilon n^{-p} \quad \text{as} \quad h \to 0,$$

for a constant C_ϵ that depends on ϵ. However, this asymptotic convergence rate is only observed for $h \lesssim \epsilon$; i.e., only after the boundary layer is resolved. In the pre-asymptotic regime, the error behaves as

$$\|u - u_h\|_{H^1(\Omega)} \sim C_\epsilon^{\text{pre}} h^{1/2} \leq C_\epsilon^{\text{pre}} n^{-1/2} \quad \text{for } h \gtrsim \epsilon.$$

Hence, while the approximation is asymptotically quasi-optimal as $h \to 0$, it exhibits slow convergence for $h \gtrsim \epsilon$.

We now consider a graded mesh whose element diameter varies in $\Omega \equiv (0,1)$ according to

$$h(x) \approx c \exp(\nu x),$$

for some $c > 0$ and a grading parameter $\nu > 0$. If the grading parameter ν is optimally chosen for a given boundary layer thickness ϵ and the polynomial degree p, then we can essentially eliminate the pre-asymptotic behavior and achieve

$$\|u - u_h\|_{H^1(\Omega)} \leq C'_\epsilon n^{-p}$$

for all n, where C'_ϵ depends only weakly on ϵ. Hence, in the case of singular perturbations, while the formal convergence rate with respect to the number of degrees of freedom is not improved by the graded meshes (since it is already optimal on uniform meshes), adaptive mesh refinement can in practice decrease the number of degrees of freedom required to achieve a given accuracy.

8.7 Summary

We summarize key points of this lecture:

1. The solution to the Poisson equation on domain with corners are not smooth in general; specifically, the solution lies in $H^{s+1}(\Omega)$ if and only if $s \in (0, \pi/\omega)$, where ω is the corner angle.

2. Adaptive finite element methods build on two technical integredients: (localizable) *a posteriori* error estimates and adaptive mesh refinement.

3. In residual error estimate, the solution error is bounded as a function of the coercivity constant and the dual norm of the residual. The coercivity constant can be estimated by solving a finite-element approximation of the eigenproblem; the residual can be estimated based on the local element and facet residuals.

4. Extrapolation error estimate allows us to estimate the error using a hierarchy of solutions for any equations and in any norms or for functional outputs.

5. Adaptive mesh refinement builds on four steps: SOLVE, ESTIMATE, MARK, and REFINE.

6. In the presence of singularities, adaptive mesh refinement can improve the formal asymptotic convergence rate (with respect to the number of degrees of freedom).

7. In the presence of a boundary layer (which is formally smooth), the adaptive mesh refinement can reduce the number of degrees of freedom required to enter the asymptotic regime.

Lecture 9

Hyperbolic and advection-dominated problems: Galerkin least-squares method

9.1 Motivation

In this lecture we consider hyperbolic and advection-dominated problems, which arise in fluid mechanics. As we will see shortly, the standard Galerkin method is ill-suited for advection-dominated problems; the approximation exhibits spurious oscillations on coarse meshes due to insufficient dissipation. In order to overcome this shortcoming, in this lecture we consider stabilized finite element methods, and in particular the Galerkin least-squares method.

9.2 Problem description

We first introduce a Lipshitz domain $\Omega \subset \mathbb{R}^d$. We then introduce an advection field $b \in L^\infty(\Omega)^d$ and identify the associated *inflow* and *outflow boundaries*:

$$\begin{aligned} \Gamma_{\text{in}} &\equiv \{x \in \partial\Omega \mid n(x) \cdot b(x) < 0\}, \\ \Gamma_{\text{out}} &\equiv \partial\Omega \setminus \Gamma_{\text{in}}. \end{aligned}$$

We next introduce a Dirichlet boundary Γ_D and a Neumann boundary Γ_N such that $\Gamma_D \cap \Gamma_N = \emptyset$ and $\overline{\partial\Omega} = \bar{\Gamma}_D \cup \bar{\Gamma}_N$. We in addition assume that the entire inflow boundary is Dirichlet: i.e., $\Gamma_{\text{in}} \subset \Gamma_D$. This condition is necessary for the equation to be well posed in the limit of vanishing diffusion. The strong form of the advection-diffusion equation is

$$\begin{aligned} -\nabla \cdot (\kappa \nabla u) + \nabla \cdot (bu) &= f \quad \text{in } \Omega, \\ u &= u^{\text{b}} \quad \text{on } \Gamma_D, \\ n \cdot \kappa \nabla u &= g \quad \text{on } \Gamma_N, \end{aligned} \tag{9.1}$$

where $\kappa \in L^\infty(\Omega)$ is the diffusion field, $b \in L^\infty(\Omega)^d$ is the advection field, $f \in L^2(\Omega)$ is the source term, $u^b \in H^1(\Gamma_D)$ is Dirichlet boundary function, and $g \in L^2(\Gamma_N)$ is the Neumann source function. We assume $\kappa(x) \geq \kappa_{\min} > 0$ a.e. in Ω. For simplicity we consider a scalar (as opposed to tensor) diffusion field and divergence-free advection field so that $\nabla \cdot b = 0$; both of these assumptions can be readily relaxed.

If $\kappa = 0$ everywhere, then the advection-diffusion equation (9.1) becomes an advection equation, which is hyperbolic. The strong from of the advection equation is

$$\begin{aligned} \nabla \cdot (bu) &= f \quad \text{in } \Omega, \\ u &= u^{\mathrm{b}} \quad \text{on } \Gamma_D = \Gamma_{\mathrm{in}}, \end{aligned} \tag{9.2}$$

with Dirichlet boundary condition on the inflow boundary Γ_{in}, and no boundary conditions on the outflow boundary Γ_{out}.

We now consider the weak form of the problem. To this end, we first introduce a Hilbert space

$$\mathcal{V} \equiv \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\} \tag{9.3}$$

and an affine space

$$\mathcal{V}^E = u^E + \mathcal{V},$$

where $u^E \in H^1(\Omega)$ is any function such that $u^E|_{\Gamma_D} = u^{\mathrm{b}}$. The space $\mathcal{V}$ is endowed with an inner product $(\cdot,\cdot)_{\mathcal{V}} \equiv (\cdot,\cdot)_{H^1(\Omega)}$ and the associated induced norm $\|\cdot\|_{\mathcal{V}} \equiv \|\cdot\|_{H^1(\Omega)}$. We then multiply the strong form of the equation by a test function $v \in \mathcal{V}$, integrate by parts, and obtain the weak formulation: find $u \in \mathcal{V}^E$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathcal{V}, \tag{9.4}$$

where

$$a(w, v) \equiv \underbrace{\int_\Omega \nabla v \cdot \kappa \nabla w dx}_{a^\kappa(w,v)} \underbrace{- \int_\Omega \nabla v \cdot b w dx + \int_{\Gamma_N} v(n \cdot b) w ds}_{a^b(w,v)} \quad \forall w, v \in \mathcal{V} \tag{9.5}$$

$$\ell(v) \equiv \int_\Omega v f dx + \int_{\Gamma_N} v g ds, \quad \forall v \in \mathcal{V}. \tag{9.6}$$

Note, unlike our previous treatment of the equation in e.g., Section 2.8, we apply integration by parts to also the advection term; the resulting bilinear form is represented in a different form but of course is identical to the previous form for functions in $\mathcal{V}$. We also identify the decomposition of the bilinear form into the forms associated with the diffusion and advection contributions, $a^\kappa(\cdot,\cdot)$ and $a^b(\cdot,\cdot)$, respectively.

9.3 Weak formulation: analysis

We now analyze the well-posedness of the advection-diffusion equation (9.4). To this end, we will first show that the bilinear form (9.5) is coercive and continuous and the linear form (9.6) is continuous in $\mathcal{V}$.

Proposition 9.1. The bilinear form (9.5) is coercive and continuous in $\mathcal{V}$ given by (9.3) with constants

$$\alpha = \frac{\kappa_{\min}}{1 + C_{\mathrm{PF}}}$$
$$\gamma = \|\kappa\|_{L^\infty(\Omega)} + \|b\|_{L^\infty(\Omega)} + C_{\mathrm{tr}}^2 \|b\|_{L^\infty(\Gamma_N)},$$

where C_{PF} is the Poincaré-Friedrichs constant such that $\|v\|^2_{L^2(\Omega)} \le C_{\mathrm{PF}} |v|^2_{H^1(\Omega)}$ $\forall v \in \mathcal{V}$, and C_{tr} is the trace inequality constant such that $\|v\|_{L^2(\Gamma_N)} \le C_{\mathrm{tr}} \|v\|_{H^1(\Omega)}$ $\forall v \in \mathcal{V}$.

Proof. We first analyze the coercivity of the bilinear form. We first note that, $\forall v \in \mathcal{V}$,

$$\begin{aligned} a^b(v,v) &= -\int_\Omega \nabla v \cdot b v dx + \int_{\Gamma_N} v(n\cdot b) v ds \\ &= -\frac{1}{2}\int_\Omega \nabla v \cdot b v dx + \frac{1}{2}\int_\Omega v b \cdot \nabla v dx - \frac{1}{2}\int_{\Gamma_N} (n\cdot b) v^2 ds + \int_{\Gamma_N} (n\cdot b) v^2 ds \\ &= \frac{1}{2}\int_{\Gamma_N} (n\cdot b) v^2 ds; \end{aligned}$$

here the first equality follows from integration by parts of $-\frac{1}{2}\int_\Omega \nabla v \cdot b v dx$ and $\nabla \cdot b = 0$. We next note that by the Poincaré-Friedrichs inequality,

$$\begin{aligned} |v|^2_{H^1(\Omega)} &= \frac{1}{1+C_{\mathrm{PF}}}|v|^2_{H^1(\Omega)} + \frac{C_{\mathrm{PF}}}{1+C_{\mathrm{PF}}}|v|^2_{H^1(\Omega)} \ge \frac{1}{1+C_{\mathrm{PF}}}|v|^2_{H^1(\Omega)} + \frac{1}{1+C_{\mathrm{PF}}}\|v\|^2_{L^2(\Omega)} \\ &= \frac{1}{1+C_{\mathrm{PF}}}\|v\|^2_{H^1(\Omega)}; \end{aligned}$$

here the second inequality follows from $\|v\|^2_{L^2(\Omega)} \le C_{\mathrm{PF}} |v|^2_{H^1(\Omega)}$. Hence it follows that

$$a(v,v) = \int_\Omega \nabla v \cdot \kappa \nabla v dx + \frac{1}{2}\int_{\Gamma_N} (n\cdot b) v^2 ds \ge \kappa_{\min} |v|^2_{H^1(\Omega)} \ge \frac{\kappa_{\min}}{1+C_{\mathrm{PF}}}\|v\|^2_{H^1(\Omega)};$$

here the first inequality follows form (i) $\kappa(x) \ge \kappa_{\min} > 0$ a.e. in Ω, and (ii) $n \cdot b \ge 0$ on $\Gamma_N \subset \Gamma_{\mathrm{out}}$.

We next analyze the continuity of the bilinear form. We have, $\forall v \in \mathcal{V}$,

$$\begin{aligned} |a(w,v)| &= \left| \int_\Omega \nabla v \cdot \kappa \nabla w dx - \int_\Omega \nabla v \cdot b w dx + \int_{\Gamma_N} v(n\cdot b) w ds \right| \\ &\le \|\kappa\|_{L^\infty(\Omega)} |v|_{H^1(\Omega)} |w|_{H^1(\Omega)} + \|b\|_{L^\infty(\Omega)} |v|_{H^1(\Omega)} |w|_{L^2(\Omega)} + \|b\|_{L^\infty(\Gamma_N)} \|v\|_{L^2(\Gamma_N)} \|w\|_{L^2(\Gamma_N)} \\ &\le \|\kappa\|_{L^\infty(\Omega)} |v|_{H^1(\Omega)} |w|_{H^1(\Omega)} + \|b\|_{L^\infty(\Omega)} |v|_{H^1(\Omega)} |w|_{L^2(\Omega)} + C_{\mathrm{tr}}^2 \|b\|_{L^\infty(\Gamma_N)} \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)} \\ &\le (\|\kappa\|_{L^\infty(\Omega)} + \|b\|_{L^\infty(\Omega)} + C_{\mathrm{tr}}^2 \|b\|_{L^\infty(\Gamma_N)}) \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)}, \end{aligned}$$

where we have invoked the trace inequality $\|v\|_{L^2(\partial\Omega)} \le \|v\|_{L^2(\Gamma_N)} \le C_{\mathrm{tr}} \|v\|_{H^1(\Omega)}$ $\forall v \in H^1(\Omega)$. □

Proposition 9.2. The linear form (9.6) is continuous in $\mathcal{V}$ given by (9.3) with a continuity constant

$$C_\ell = \|f\|_{L^2(\Omega)} + C_{\mathrm{tr}} \|g\|_{L^2(\Gamma_N)}.$$

Proof. We observe that, $\forall v \in \mathcal{V}$,

$$|\ell(v)| = \left| \int_\Omega v f dx + \int_{\Gamma_N} v g ds \right| \le \|v\|_{L^2(\Omega)} \|f\|_{L^2(\Omega)} + \|v\|_{L^2(\Gamma_N)} \|g\|_{L^2(\Gamma_N)}$$
$$\le \|v\|_{L^2(\Omega)} \|f\|_{L^2(\Omega)} + C_{\rm tr} \|v\|_{H^1(\Omega)} \|g\|_{L^2(\Gamma_N)} \le (\|f\|_{L^2(\Omega)} + C_{\rm tr} \|g\|_{L^2(\Gamma_N)}) \|v\|_{H^1(\Omega)},$$

which is the desired result. □

Proposition 9.3. The advection-diffusion equation (9.4) has a unique solution.

Proof. Because the bilinear form $a(\cdot,\cdot)$ is coercive and continuous and the linear form $\ell(\cdot)$ is continuous, the existence and uniqueness of the solution follows from the Lax-Milgram theorem. □

9.4 Standard Galerkin method: limitations

Because the bilinear form is coercive and continuous and the linear form is continuous, we can readily consider the (standard) Galerkin approximation of the weak formulation (9.4). To this end, we introduce a $\mathbb{P}^p$ approximation space

$$\mathcal{V}_h \equiv \{v \in \mathcal{V} \mid v|_K \in \mathbb{P}^p(K),\ \forall K \in \mathcal{T}_h\}$$

associated with the triangulation $\mathcal{T}_h$. We also introduce the affine space $\mathcal{V}_h^E = u^E + \mathcal{V}_h$. We then consider the following finite element approximation: find $u_h \in \mathcal{V}_h^E$ such that

$$a(u_h, v) = \ell(v) \quad \forall v \in \mathcal{V}_h. \tag{9.7}$$

Because $\mathcal{V}_h \subset \mathcal{V}$, the conditions of the Lax-Milgram theorem also holds for $\mathcal{V}_h$, and (9.7) is wellposed.

In practice, however, this finite element approximation exhibits unsatisfactory oscillations when the ratio of the diffusion to advection is small with respect to the element diamester in the sense that grid Péclet number $\mathrm{Pe}_h \equiv \frac{|b| h_K}{2\kappa}$ is much greater than unity. An example of such a failure is shown in Figure 9.1. Here the advection-diffusion equation with $\kappa = 1/50$ and $b = 1$ is solved using the Galerkin finite element method for $h = 1/8$; the grid Péclet number is $\mathrm{Pe}_h = 12.5 \gg 1$. This poor behavior of the Galerkin method for advection-dominated problems does not violate the quasi-optimality result provided by the Céa's lemma:

$$\|u - u_h\|_{\mathcal{V}} \le \frac{\gamma}{\alpha} \inf_{w_h \in \mathcal{V}_h} \|u - w_h\|_{\mathcal{V}}.$$

The gap between the $\mathcal{V}$-optimal approximation and the finite element approximation for advection-dominated problems is $\gamma/\alpha = \|b\|_{L^\infty(\Omega)}/\kappa_{\min} \gg 1$; while we do not necessarily expect the method to perform poorly as the lemma only provides an upper bound, we do not have an *a priori* guarantee that the method works well. This contrasts with, say, Poisson's equation, for which we know the Galerkin method provides a quasi-optimal solution with a small gap to the best-fit solution.

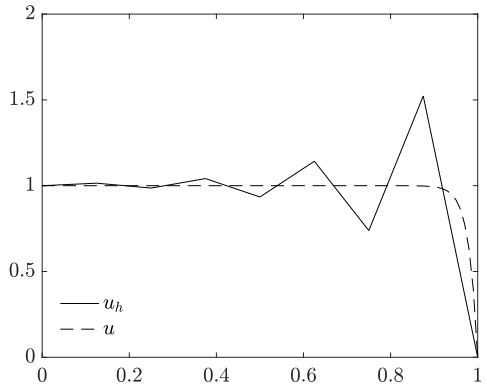


Figure 9.1: Failure of the Galerkin method for one-dimensional advection diffusion with $\kappa = 1/50$, $b = 1$, and $h = 1/8$.

9.5 Artificial diffusion method

The instability of the standard Galerkin method for $\mathrm{Pe}_h = \frac{|b|h_K}{2\kappa} \gg 1$ is due to the insufficient amount of "apparent" diffusion provided in the boundary layers when the features are underresolved. One way to overcome the instability hence is to artificially add diffusion that scales with h, such that Pe_h of the modified equation is of order unity. The approach that simply modifies the diffusion coefficient κ to scale with h is called the *artificial diffusion method* The modified bilinear form is of the form

$$a_h(w, v) = \int_\Omega \nabla v \cdot (\kappa + ch)\nabla w dx - \int_\Omega \nabla v \cdot bw dx + \int_{\Gamma_N} v(n \cdot b)w ds,$$

for $c \approx 1$. However, this approach has two major limitations. First, the artificial diffusion is applied isotropically in all directions, and in particular also in the direction perpendicular to the streamlines; as a result, the method excessively diffuses shear layers. Second, due to the $\mathcal{O}(h)$ modification of the bilinear form, the method is at most first-order accurate, even if the underlying solution is smooth and a higher-order approximation is used. Due to these two limitations, the artificial diffusion method is not a recommended approach to stabilize the standard Galerkin method.

9.6 Galerkin least-squares method: formulation

To overcome the instability of the standard Galerkin method, we introduce the *Galerkin least-squares (GLS) method*. By way of preliminaries, we introduce the differential operator associated with the strong form:

$$\mathcal{L}w \equiv -\nabla \cdot (\kappa \nabla w) + \nabla \cdot (bw). \tag{9.8}$$

The GLS method is a stabilized method, which employs a h-dependent forms $a_h(\cdot, \cdot)$ and $\ell_h(\cdot)$ that are different from the original forms $a(\cdot, \cdot)$ and $\ell(\cdot)$ given by (9.5) and (9.6), respectively.

Specifically, the forms are augmented by the terms arising from "least-squares" stabilization:

$$\begin{aligned} a_h(w,v) &\equiv \underbrace{a(w,v)}_{\text{Galerkin}} + \underbrace{(\tau\mathcal{L}v, \mathcal{L}w)_{L^2(\Omega)}}_{\text{least-squares stabilization}} \\ &= \int_\Omega \nabla v \cdot \kappa \nabla v dx - \int_\Omega \nabla v \cdot b w dx + \int_{\Gamma_N} v(n\cdot b) w ds + \int_\Omega (\mathcal{L}v)\tau(\mathcal{L}w) dx, \\ \ell_h(v) &\equiv \ell(v) + (\tau\mathcal{L}v, f)_{L^2(\Omega)} = \int_\Omega v f dx + \int_{\Gamma_N} v g ds + \int_\Omega (\mathcal{L}v)\tau f dx, \end{aligned}$$

where τ is the GLS stabilization parameter given by

$$\tau = \left(\left(\frac{2|b|}{h_K}\right)^2 + 9\left(\frac{4\kappa}{h_K^2}\right)^2 \right)^{-1/2}.$$

The GLS problem is as follows: find $u_h \in \mathcal{V}_h^E$ such that

$$a_h(u_h, v) = \ell_h(v) \quad \forall v \in \mathcal{V}_h. \tag{9.9}$$

We make a few remarks. First, there are other choices of the stabilization parameter τ, but in order obtain stability for $\text{Pe}_h \equiv \frac{|b|h_K}{2\kappa} \gg 1$ and convergence for $\text{Pe}_h \ll 1$, the parameter must satisfy

$$\tau = \begin{cases} \mathcal{O}\left(\frac{h_K}{|b|}\right), & \text{Pe}_h \gg 1 \\ \mathcal{O}\left(\frac{h_K^2}{\kappa}\right), & \text{Pe}_h \ll 1 \end{cases}.$$

Second, unlike the forms we have seen so far, the linear and bilinear forms of the GLS method are h-dependent because τ depends on h; we hence denote the bilinear and linear forms with a subscript h.

9.7 Galerkin least-squares method: analysis

We now analyze the GLS method. For simplicity, throughout this analysis we assume that the diffusion field κ and the advection field b are constant over Ω. We first note that the least-squares term $(\tau\mathcal{L}v, \mathcal{L}w)_{L^2(\Omega)}$ is non-negative and hence

$$a_h(v,v) \geq a(v,v) \geq \alpha \|v\|_{H^1(\Omega)} \quad \forall v \in \mathcal{V}_h;$$

the GLS bilinear form is coercive in $\mathcal{V}_h$. Assuming an inverse estimate $|v|_{H^2(\kappa)} \leq ch^{-1}\|v\|_{H^1(\kappa)}$ $\forall v \in \mathbb{P}^p(\kappa)$ holds, we can readily show that the GLS bilinear and linear forms are also continuous in $\mathcal{V}_h$. It hence follows that GLS problem (9.9) has a unique solution by the Lax-Milgram theorem.

We second note that the least-squares term $(\tau\mathcal{L}v, \mathcal{L}w)_{L^2(\Omega)}$ provides diffusion in the streamline direction. To see this, we consider $\kappa = 0$ for simplicity and observe that

$$(\tau\mathcal{L}v, \mathcal{L}w)_{L^2(\Omega)} = \int_\Omega \frac{h_K}{2|b|}(\nabla\cdot(bv))(\nabla\cdot(bw))dx = \int_\Omega \frac{h_K}{2|b|}(b\cdot\nabla v)(b\cdot\nabla w)dx,$$

where the last equality follows form the assumption $\nabla\cdot b = 0$. We observe that the (i) the artificial diffusion is added in the direction of b and (ii) the strength of the diffusion is $\mathcal{O}(h_K|b|)$. This streamline diffusion adds the requires stability to the Galerkin formulation.

However, as discussed in the context of the (classical) artificial diffusion method, if we simply make an $\mathcal{O}(h)$ modification to the bilinear form, then the resulting method will be at most first order. To overcome this limitation, the GLS method modifies the standard Galerkin method such that the modified forms are *consistent*. A stabilized method (or more generally a method that uses $a(\cdot,\cdot)$ and $\ell(\cdot)$ that are different from the original weak formulation) is said to be consistent if the following holds.

Definition 9.4 (consistency)**.** Suppose the exact solution $u \in \mathcal{V}$ that satisfies $a(u,v) = \ell(v)\ \forall v \in \mathcal{V}$ is sufficiently smooth. A stabilized method is said to be consistent if for this smooth $u \in \mathcal{V}$

$$a_h(u,v) = \ell_h(v) \quad \forall v \in \mathcal{V}_h.$$

In other words, the exact solution satisfies the weak statement associated with the stabilized problem.

Proposition 9.5. The GLS method is consistent.

Proof. The proof follows from the definition of the stabilized forms:

$$\begin{aligned} r_h(v) &\equiv \ell_h(v) - a_h(u,v) \\ &= \ell(v) + (\tau\mathcal{L}v, f)_{L^2(\Omega)} - a(u,v) - (\tau\mathcal{L}v, \mathcal{L}u)_{L^2(\Omega)} \\ &= \underbrace{\ell(v) - a(u,v)}_{\substack{=0 \text{ since } a(u,v)=\ell(v) \\ \forall v \in \mathcal{V}_h \subset \mathcal{V}}} + (\tau\mathcal{L}v, \underbrace{f - \mathcal{L}u}_{=0 \text{ by strong form}})_{L^2(\Omega)} = 0 \quad \forall v \in \mathcal{V}_h, \end{aligned}$$

which is the desired result. □

Remark 9.6. If a stabilized method is consistent, then Galerkin orthogonality holds with respect to $a_h(\cdot,\cdot)$:

$$a_h(u - u_h, v) = \ell_h(v) - a_h(u_h, v) = 0 \quad \forall v \in \mathcal{V}_h.$$

The Galerkin orthogonality, as usual, plays a key role in our error analysis.

Thanks to the consistency (and Galerkin orthogonality) of the GLS formulation, we can show that the GLS can achieve higher-order accuracy if the underlying solution is smooth and a higher-degree polynomials are used. Specifically, we may assess the error in the GLS method in its natural norm.

Lemma 9.7. Let $\|\cdot\|_{a_h} : \mathcal{V} \to \mathbb{R}$ be the GLS norm given by

$$\|v\|_{a_h}^2 \equiv a_h(v,v) \quad \forall v \in \mathcal{V}.$$

The error in the GLS approximation is bounded by

$$\|e\|_{a_h}^2 \leq 2\|\tau^{-1/2}(u - \mathcal{I}_h u)\|_{L^2(\Omega)}^2 + 2\|\tau^{1/2}\mathcal{L}(u - \mathcal{I}_h u)\|_{L^2(\Omega)}^2,$$

where $\mathcal{I}_h : \mathcal{V} \to \mathcal{V}_h$ is the interpolation operator.

Proof. Let $\eta_h \equiv u - \mathcal{I}_h u$, and we observe that

$$\begin{aligned}\|e\|_{a_h}^2 &= a_h(e,e) = a_h(e, u - \mathcal{I}_h u) - a_h(e, \mathcal{I}_h u - u_h) = a_h(e, \eta_h) \\ &= (\eta_h, \mathcal{L}e)_{L^2(\Omega)} + (\tau \mathcal{L}\eta_h, \mathcal{L}e)_{L^2(\Omega)} \\ &\leq \|\tau^{-1/2}\eta_h\|_{L^2(\Omega)}^2 + \frac{1}{4}\|\tau^{1/2}\mathcal{L}e\|_{L^2(\Omega)}^2 + \|\tau^{1/2}\mathcal{L}\eta_h\|_{L^2(\Omega)}^2 + \frac{1}{4}\|\tau^{1/2}\mathcal{L}e\|_{L^2(\Omega)}^2 \\ &\leq \frac{1}{2}\|e\|_{a_h}^2 + \|\tau^{-1/2}\eta_h\|_{L^2(\Omega)}^2 + \|\tau^{1/2}\mathcal{L}\eta_h\|_{L^2(\Omega)}^2. \end{aligned} \tag{9.10}$$

A few elaborations are in order. The third equality follows from the Galerkin orthogonality and the definition $\eta_h = u - \mathcal{I}_h u$. The first inequality relies on the so-called algebraic-mean geometric-mean inequality: $\forall a, b \in \mathbb{R}$ and $\gamma > 0$, $|ab| \leq \frac{1}{2\gamma}a^2 + \frac{\gamma}{2}b^2$; using the inequality we obtain

$$\begin{aligned}(\eta_h, \mathcal{L}e)_{L^2(\Omega)} &\leq \|\tau^{-1/2}\eta_h\|_{L^2(\Omega)}^2 + \frac{1}{4}\|\tau^{1/2}\mathcal{L}e\|_{L^2(\Omega)}^2, \\ (\tau\mathcal{L}\eta_h, \mathcal{L}e)_{L^2(\Omega)} &\leq \|\tau^{1/2}\mathcal{L}\eta_h\|_{L^2(\Omega)}^2 + \frac{1}{4}\|\tau^{1/2}\mathcal{L}e\|_{L^2(\Omega)}^2. \end{aligned}$$

The second inequality follows from the definition of $\|\cdot\|_{a_h}^2$. □

We now specialize Lemma 9.7 to two separate cases. First we consider the advection equation $\kappa = 0$.

Proposition 9.8. Consider the GLS approximation of the advection (9.1) for which $\kappa = 0$. If $u \in H^1(\Omega) \cap H^{s+1}(\mathcal{T}_h)$, the error in the GLS approximation is in the GLS norm by

$$\|u - u_h\|_{a_h} \leq Ch^{r+1/2}|u|_{H^{r+1}(\mathcal{T}_h)}$$

for $r = \min\{s, p\}$, and in particular the error in the streamwise derivative is bounded by

$$\|b \cdot \nabla(u - u_h)\|_{L^2(\Omega)} \leq \tilde{C}h^r|u|_{H^{r+1}(\mathcal{T}_h)}.$$

Proof. We first note that for $\kappa = 0$, $\tau = h/(2\|b\|)$. We then simplify Lemma 9.7 as

$$\begin{aligned}\|e\|_{a_h}^2 &\leq 2\|\tau^{-1/2}\eta_h\|_{L^2(\Omega)}^2 + 2\|\tau^{1/2}\mathcal{L}\eta_h\|_{L^2(\Omega)}^2 = 2\|\tau^{-1/2}\eta_h\|_{L^2(\Omega)}^2 + 2\|\tau^{1/2}b\cdot\nabla\eta_h\|_{L^2(\Omega)}^2 \\ &\leq \frac{4\|b\|}{h}\|\eta_h\|_{L^2(\Omega)}^2 + h\|b\|\|\nabla\eta_h\|_{L^2(\Omega)}^2 \leq 4\|b\|h^{-1}C_{\mathcal{I}}h^{2r+2}|u|_{H^{r+1}(\mathcal{T}_h)}^2 + h\|b\|C_{\mathcal{I}}'h^{2r}|u|_{H^{r+1}(\mathcal{T}_h)}^2 \\ &\leq Ch^{2r+1}|u|_{H^{r+1}(\mathcal{T}_h)}^2; \end{aligned}$$

here the first inequality follows from Lemma 9.7, the first equality follows from $\mathcal{L}(\cdot) = b \cdot \nabla(\cdot)$, the second inequality follows from the definition of τ, the third inequality follows from the interpolation error bounds. In addition, we note that for $\kappa = 0$,

$$\begin{aligned}\|v\|_{a_h}^2 \equiv a_h(v,v) &= \frac{1}{2}\int_{\Gamma_N}(n\cdot b)v^2 ds + \int_\Omega (\mathcal{L}v)\tau(\mathcal{L}v)dx = \frac{1}{2}\|(n\cdot b)^{1/2}v\|_{L^2(\Gamma_N)}^2 + \|\tau^{1/2}\mathcal{L}v\|_{L^2(\Omega)}^2 \\ &= \frac{1}{2}\|(n\cdot b)^{1/2}v\|_{L^2(\Gamma_N)}^2 + \frac{h}{2|b|}\|b\cdot\nabla v\|_{L^2(\Omega)}^2. \end{aligned}$$

Hence

$$\|b\cdot\nabla e\|_{L^2(\Omega)}^2 \leq 2\|b\|_2 h^{-1}\|e\|_{a_h}^2 \leq Ch^{2r}|u|_{H^{r+1}(\mathcal{T}_h)}^2,$$

which is the desired second inequality. □

We make a few observations. We first observe that the GLS stabilization provides the control of the error in the derivative in the streamline direction, and we obtain the optimal order of h^r. We second observe that, as expected for a purely hyperbolic equations, we have no control of the error in the derivatives in the directions normal to the streamline directions.

We now specialize Lemma 9.7 for $\kappa > 0$.

Proposition 9.9. Consider the GLS approximation of the advection-diffusion (9.1) for which $\kappa_{\min} > 0$. If $u \in H^2(\Omega) \cap H^{s+1}(\mathcal{T}_h)$, the error in the GLS solution is bounded by

$$\|u - u_h\|_{a_h} \leq Ch^r |u|_{H^{r+1}(\mathcal{T}_h)}$$

for $r = \min\{s, p\}$. Moreover, because the $\|\cdot\|_{a_h}$ is equivalent to $\|\cdot\|_{H^1(\Omega)}$ for $\kappa_{\min} > 0$, it follows that

$$\|u - u_h\|_{H^1(\Omega)} \leq \tilde{C}h^r |u|_{H^{r+1}(\mathcal{T}_h)}.$$

Proof. We first note that for $\kappa > 0$, $\tau = \mathcal{O}(h^2)$ as $h \to 0$. We then simplify Lemma 9.7 as

$$\begin{aligned}\|e\|_{a_h}^2 &\leq 2\|\tau^{-1/2}\eta_h\|_{L^2(\Omega)}^2 + 2\|\tau^{1/2}\mathcal{L}\eta_h\|_{L^2(\Omega)}^2 \leq ch^{-2}\|\eta_h\|_{L^2(\Omega)}^2 + c'h^2\|\Delta\eta_h\|_{L^2(\Omega)}^2 \\ &\leq ch^{-2}C_{\mathcal{I}}h^{2r+2}|u|_{H^{r+1}(\mathcal{T}_h)}^2 + c'h^2C_{\mathcal{I}}h^{2r-2}|u|_{H^{r+1}(\mathcal{T}_h)}^2 \leq Ch^{2r}|u|_{H^{r+1}(\mathcal{T}_h)}^2;\end{aligned}$$

here the first inequality follows from Lemma 9.7, the second inequality follows from the definition of τ as $h \to 0$, and the third inequality follows from interpolation error bounds. □

We make a few remarks. We first note that, in the presence of diffusion as $h \to 0$, the GLS formulation recovers the optimal convergence rate of h^r in the $H^1(\Omega)$ norm; this is unlike the (classical) artificial diffusion formulation, which is at most first order. We second remark that the asymptotic error analysis cannot highlight the additional stability provided by the GLS formulation, because the stabilization in fact vanishes as $h \to 0$; however, in practice the GLS method, unlike the standard Galerkin method, yields a stable approximation when $\mathrm{Pe}_h \gg 1$.

9.8 Streamline-upwind Petrov-Galerkin (SUPG) method

The GLS method is closely related to the streamline-upwind Petrov-Galerkin (SUPG) method, which precedes the GLS method. The bilinear and linear forms for the SUPG method are given by

$$\begin{aligned}a_h(w, v) &\equiv a(w, v) + (\tau b \cdot \nabla v, \mathcal{L}w)_{L^2(\Omega)} \\ &= \int_\Omega \nabla v \cdot \kappa \nabla v dx - \int_\Omega \nabla v \cdot bw dx + \int_{\Gamma_N} v(n \cdot b)w ds + \int_\Omega (b \cdot \nabla v)\tau(\mathcal{L}w)dx, \\ \ell_h(v) &\equiv \ell(v) + (\tau b \cdot \nabla v, f)_{L^2(\Omega)} = \int_\Omega vf dx + \int_\Omega (b \cdot \nabla v)\tau f dx.\end{aligned}$$

In other words, the SUPG method is obtained by replacing the least-squares stabilization term $(\tau\mathcal{L}v, \mathcal{L}w - f)_{L^2(\Omega)}$ with $(\tau b \cdot \nabla v, \mathcal{L}w - f)_{L^2(\Omega)}$.

The SUPG method is called a *Petrov-Galerkin method*, because the method can be interpreted as using modified test functions. Specifically, if $\Gamma_N = \emptyset$ (i.e., $\partial\Omega = \Gamma_D$), then

$$\begin{aligned}a_h(w, v) &= (v, \mathcal{L}w)_{L^2(\Omega)} + (\tau b \cdot \nabla v, \mathcal{L}w)_{L^2(\Omega)} = (v + \tau b \cdot \nabla v, \mathcal{L}w)_{L^2(\Omega)}, \\ \ell_h(w, v) &= (v, f)_{L^2(\Omega)} + (\tau b \cdot \nabla v, f)_{L^2(\Omega)} = (v + \tau b \cdot \nabla v, f)_{L^2(\Omega)}.\end{aligned}$$

Hence the SUPG formulation may be interpreted as a weighted-residual method with test functions $v + \tau b \cdot \nabla v$. These test functions may be considered "upwinded", as they place more weight to the upwind side.

We make a few additional remarks. First, the SUPG method, like the GLS method, adds artificial diffusion in the direction of the streamlines and hence provides additional stability for advection-dominated problems. Second, the SUPG method, like the GLS method, is consistent. Third, the SUPG and GLS methods are identical in some cases: (i) for advection equations (i.e., $\kappa = 0$) the methods are identical because $\mathcal{L}v = b \cdot \nabla v$; (ii) for $\mathbb{P}^1$ approximations, the methods are identical because $\mathcal{L}v_h = -\kappa\Delta v_h + b \cdot \nabla v_h = b \cdot \nabla v_h$.

9.9 Summary

We summarize key points of this lecture:

1. The weak formulation of advection-diffusion equation yields a coercive and continuous bilinear form and continuous linear form and hence is well-posed.

2. For advection-dominated problems, the Galerkin finite element approximation exhibits spurious oscillation in boundary layers when the grid Péclet number $\mathrm{Pe}_h \equiv \frac{|b|h_K}{2\kappa} \gg 1$.

3. The GLS method adds a least-squares term to the standard Galerkin method to stabilize the approximation. This stabilization adds diffusion in the streamline direction and is consistent. The stabilization removes the spurious oscillation in underresolved boundary layers.

4. The GLS method provides control of the error in the streamline derivative even in the hyperbolic limit of no diffusion. The error is optimal (i.e., h^p) for the streamline derivative, and suboptimal (i.e., $h^{p+1/2}$) in the L^2 norm; however, in practice we often observe the optimal convergence rate of h^{p+1} in the $L^2(\Omega)$ norm for smooth problems.

5. The GLS method provides optimal convergence rate of h^p in the $H^1(\Omega)$ norm for advection-diffusion equations with a smooth solution.

6. The SUPG method is closely related the GLS method; the SUPG method also adds artificial diffusion in the streamline direction in a consistent manner, and the method can be interpreted as a Petrov-Galerkin method.

Lecture 10

Parabolic equations

10.1 Motivation

We have so far considered stationary PDEs; however, many phenomena in continuum mechanics are time-dependent and hence are modeled by time-dependent PDEs. In this lecture we consider variational formulations, and the associated finite element approximation, of time-dependent PDEs. Of particular focus in the lecture is parabolic PDEs — PDEs that are first-order in time and whose spatial operator is elliptic; the heat equation, which describes unsteady heat transfer, is a prototypical parabolic PDE.

10.2 Model equation: heat equation

We first introduce our model parabolic equation: the heat equation. Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain and $\mathcal{I} \equiv (0,T]$ be the time interval. We partition the boundary $\partial\Omega$ into a Dirichelt boundary Γ_D and a Neumann boundary Γ_N such that $\overline{\partial\Omega} = \bar{\Gamma}_D \cup \bar{\Gamma}_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$. Unlike the steady state case, we may consider $\Gamma_D = \emptyset$.

We introduce the strong form of the heat equation: find u such that

$$\begin{aligned} \frac{\partial u}{\partial t} - \nabla \cdot (\kappa \nabla u) &= f \quad \text{in } \Omega \times \mathcal{I}, \\ u &= u^{\mathrm{b}} \quad \text{on } \Gamma_D \times \mathcal{I}, \\ n \cdot \kappa \nabla u &= g \quad \text{on } \Gamma_N \times \mathcal{I}, \\ u(t=0) &= u^0 \quad \text{in } \Omega, \end{aligned}$$

where $\kappa \in L^\infty(\Omega)^{d\times d}$ is the thermal diffusivity tensor field, $f : \Omega \times \mathcal{I} \to \mathbb{R}$ is the volume heat source, $g : \Gamma_N \times \mathcal{I} \to \mathbb{R}$ is the boundary heat flux on Γ_N, $u^{\mathrm{b}} : \Gamma_D \to \mathbb{R}$ is the prescribed temperature on Γ_D, and $u^0 : \Omega \to \mathbb{R}$ is the initial condition. We assume that the spatial operator is elliptic: i.e., $\xi^T \kappa(x,t) \xi > 0 \ \forall \xi \in \mathbb{R}^d$, $\xi \neq 0$, a.e. $x \in \Omega$ and $t \in \mathcal{I}$. For simplicity, we assume that the prescribed temperature on Γ_D is independent of time. This is a *parabolic equation* because (i) the equation is first-order in time and (ii) the spatial operator is elliptic.

10.3 Variational formulation

We now derive a (spatially) weak form of the heat equation. To this end, we introduce a Hilbert space

$$\mathcal{V} \equiv \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}$$

and an affine space

$$\mathcal{V}^E \equiv u^E + \mathcal{V},$$

where u^E is any function in $H^1(\Omega)$ such that $u^E|_{\Gamma_D} = u^{\mathrm{b}}$. As in the steady-state problem, the Dirichlet conditions are essential boundary conditions that must be enforced explicitly through the choice of the space.

We next take an arbitrary test function $v \in \mathcal{V}$, multiply the governing equation by v, integrate by parts, and make appropriate substitutions for the natural boundary conditions to obtain

$$\int_\Omega v \frac{\partial u}{\partial t} dx + \int_\Omega \nabla v \cdot \kappa \nabla u dx = \int_\Omega v f dx + \int_{\Gamma_N} v g ds.$$

Our weak formulation is as follows: find $u(t) \in \mathcal{V}^E$, $t \in \mathcal{I}$, such that

$$\begin{aligned} m\left(\left.\frac{\partial u}{\partial t}\right|_t, v\right) + a(u(t), v; t) &= \ell(v; t) \quad \forall v \in \mathcal{V}, \forall t \in \mathcal{I}, \\ (u(t=0), v)_{L^2(\Omega)} &= (u^0, v)_{L^2(\Omega)} \quad \forall v \in \mathcal{V}, \end{aligned} \tag{10.1}$$

where

$$\begin{aligned} m(w, v) &\equiv \int_\Omega v w dx \quad \forall w, v \in \mathcal{V}, \\ a(w, v; t) &\equiv \int_\Omega \nabla v \cdot \kappa \nabla w dx \quad \forall w, v \in \mathcal{V}, \\ \ell(v; t) &\equiv \int_\Omega v f dx + \int_{\Gamma_N} v g ds \quad \forall v \in \mathcal{V}, \end{aligned}$$

where $f \in L^2(\Omega \times \mathcal{I})$, $g \in L^2(\Gamma_N \times \mathcal{I})$, and $u_0 \in L^2(\Omega)$. Note that the parameter t of the forms $a(\cdot, \cdot; t)$ and $\ell(\cdot; t)$ signifies the forms in general are time dependent. We note that it is also possible to consider a space-time weak formulation, where we also integrate in the time domain; we do not consider the formulation here.

Remark 10.1. While the particular forms of (10.1) are associated with the heat equation, both the formulations and analyses in the remainder of this lecture apply to general parabolic equations of the form (10.1) with a uniformly coercive and continuous $a(\cdot, \cdot; t)$ and a uniformly continuous $\ell(\cdot)$; i.e., there exist $\alpha > 0$, $\gamma < \infty$, and c_ℓ such that

$$\begin{aligned} a(v, v; t) &\geq \alpha \|v\|_{\mathcal{V}}^2 \quad \forall v \in \mathcal{V}, \forall t \in \mathcal{I}, \\ a(w, v; t) &\leq \gamma \|w\|_{\mathcal{V}} \|v\|_{\mathcal{V}} \quad \forall w, v \in \mathcal{V}, \forall t \in \mathcal{I}, \\ \ell(v; t) &\leq c_\ell \|v\|_{\mathcal{V}} \quad \forall v \in \mathcal{V}, \forall t \in \mathcal{I}. \end{aligned}$$

The parabolic equation (10.1) has a unique solution:

Proposition 10.2. Suppose that the bilinear form $a(\cdot,\cdot;t)$ is $\mathcal{V}$-coercive and $\mathcal{V}$-continuous a.e. $t \in \mathcal{I}$. Then, given $f \in L^2(\Omega \times \mathcal{I})$, $g \in L^2(\Gamma_N \times \mathcal{I})$, and $u_0 \in L^2(\Omega)$, there exists a unique solution $u \in L^2(\mathcal{I};\mathcal{V}) \cap C^0(\bar{\mathcal{I}}; L^2(\Omega))$ to (10.1).

Proof. Proof is beyond the scope of this course. We refer to Quarteroni and Valli (2008). □

Proposition 10.3. Under the assumptions of Proposition 10.2, the energy estimate

$$\|u(t)\|^2_{L^2(\Omega)} + \alpha \int_{\tau=0}^{t} \|u(\tau)\|^2_{\mathcal{V}} d\tau \leq \|u^0\|^2_{L^2(\Omega)} + \frac{1}{\alpha} \int_{\tau=0}^{t} \|\ell(\cdot;\tau)\|^2_{\mathcal{V}'} d\tau$$

holds for each $t \in \mathcal{I}$, where α is the coercivity constant.

Proof. We set $v = u(t)$ in (10.1) to obtain

$$(\left.\frac{\partial u}{\partial t}\right|_t, u(t))_{L^2(\Omega)} + a(u(t), u(t)) = \ell(u(t);t).$$

We now make three observations. First, the first term can be written as $(\left.\frac{\partial u}{\partial t}\right|_t, u(t))_{L^2(\Omega)} = \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)}$. Second, the coercivity statement yields for the second term: $a(u(t),u(t);t) \geq \alpha\|u(t)\|^2_{\mathcal{V}}$. Third, the definition of the dual norm and the Young's inequality yield $|\ell(u(t);t)| \leq \|\ell(\cdot;t)\|_{\mathcal{V}'}\|u(t)\|_{\mathcal{V}} \leq \frac{1}{2\alpha}\|\ell(\cdot;t)\|^2_{\mathcal{V}'} + \frac{\alpha}{2}\|u(t)\|^2_{\mathcal{V}}$. We hence obtain

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + \alpha\|u(t)\|^2_{\mathcal{V}} \leq \frac{1}{2\alpha}\|\ell(\cdot;t)\|^2_{\mathcal{V}'} + \frac{\alpha}{2}\|u(t)\|^2_{\mathcal{V}},$$

which further simplifies to

$$\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + \alpha\|u(t)\|^2_{\mathcal{V}} \leq \frac{1}{\alpha}\|\ell(\cdot;t)\|^2_{\mathcal{V}'}.$$

The integration of the relationship over $(0,t)$ for $t \in \mathcal{I}$ yields the desired energy statement. □

Proposition 10.3 shows that the energy at any time $t \in \mathcal{I}$ is bounded by (i) the energy at the initial time $\|u^0\|^2_{L^2(\Omega)}$ and (ii) the data ℓ, which can be further divided into the volume source term f and the boundary term g. Note that the coercive bilinear form $a(\cdot,\cdot;t)$ provides the dissipation. For a homogeneous system (i.e., $\ell = 0$), we have $\|u(t)\|^2_{L^2(\Omega)} \leq \|u^0\|^2_{L^2(\Omega)}$, and the energy strictly decays over time.

10.4 Semi-discrete formulation

Our strategy to numerical approximation of the initial value problem comprises two steps: the spatial approximation by a finite element method; the temporal approximation by a time-marching scheme for a system of ordinary differential equations (ODEs). We first consider the spatial approximation (only) to introduce a *semi-discrete form.* (The form is called semi-discrete as it is only discretized in space.) To this end, we introduce a finite element space,

$$\mathcal{V}_h = \{v \in \mathcal{V} \mid v|_K \in \mathbb{P}^p(K),\ \forall K \in \mathcal{T}_h\}.$$

We then consider the following semi-discrete problem: find $u(t) \in \mathcal{V}_h$, $t \in \mathcal{I}$, such that

$$m(\left.\frac{\partial u_h}{\partial t}\right|_t, v) + a(u_h(t), v; t) = \ell(v; t) \quad \forall v \in \mathcal{V}_h,\ t \in \mathcal{I}, \tag{10.2}$$
$$(u_h(t=0), v)_{L^2(\Omega)} = (u^0, v)_{L^2(\Omega)} \quad \forall v \in \mathcal{V}_h.$$

Since $\mathcal{V}_h \subset \mathcal{V}$, we appeal to Propositions 10.2 and 10.3 to conclude that (10.2) has a unique solution and is energy stable, respectively.

To facilitate the presentation of the (fully) discrete equation, we now introduce the algebraic form of (10.2). We first introduce a basis $\{\phi_i\}_{i=1}^n$ of the space $\mathcal{V}_h$. We then represent the solution as

$$u_h(x,t) = \sum_{j=1}^n \hat{u}_{h,j}(t)\phi_j(x),$$

where $\hat{u}_h \in \mathbb{R}^n$ is finite element coefficients; note that, because the basis is independent of time, the time derivative of the finite element solution is given by

$$\left.\frac{\partial u_h}{\partial t}\right|_{(x,t)} = \sum_{j=1}^n \left.\frac{d\hat{u}_{h,j}}{dt}\right|_t \phi_j(x).$$

We now rewrite (10.2) as follows: find $\hat{u}_h(t) \in \mathbb{R}^n$, $t \in \mathcal{I}$, such that

$$m(\sum_{j=1}^n \left.\frac{d\hat{u}_{h,j}}{dt}\right|_t \phi_j, \phi_i) + a(\sum_{j=1}^n \hat{u}_{h,j}(t)\phi_j, \phi_i; t) = \ell(\phi_i; t) \quad \forall i = 1, \ldots, n,\ t \in \mathcal{I}.$$

We then appeal to the bilinearity of $m(\cdot,\cdot)$ and $a(\cdot,\cdot;t)$ to obtain the following system of ODEs: find $\hat{u}_h(t) \in \mathbb{R}^n$, $t \in \mathcal{I}$, such that

$$\hat{M}_h \left.\frac{d\hat{u}_h(t)}{dt}\right|_t + \hat{A}_h(t)\hat{u}_h(t) = \hat{f}_h(t) \quad \text{in } \mathbb{R}^n,\ \forall t \in \mathcal{I}, \tag{10.3}$$
$$\hat{u}_h(t=0) = \hat{u}^0 \quad \text{in } \mathbb{R}^n,$$

where the mass matrix $\hat{M}_h \in \mathbb{R}^{n\times n}$, stiffness matrix $\hat{A}_h(t) \in \mathbb{R}^{n\times n}$, and load vector $\hat{f}_h(t) \in \mathbb{R}^n$ are given by

$$\hat{M}_{h,ij} = m(\phi_j, \phi_i), \quad i,j = 1, \ldots, n,$$
$$\hat{A}_{h,ij}(t) = a(\phi_j, \phi_i; t), \quad i,j = 1, \ldots, n,$$
$$\hat{f}_{h,i}(t) = \ell(\phi_i; t), \quad i = 1, \ldots, n,$$

and the initial condition vector $\hat{h} \in \mathbb{R}^n$ satisfies

$$\hat{M}_h \hat{u}^0 = \hat{h} \quad \text{in } \mathbb{R}^n$$

for $\hat{h}_i = (u^0, \phi_i)_{L^2(\Omega)}$, $i = 1, \ldots, n$. As the ODE (10.3) is simply an algebraic reformulation of the semi-discrete form (10.2) for a particular basis $\{\phi_i\}_{i=1}^n$, the ODE has a unique solution.

10.5 Semi-discrete formulation: error analysis

We now analyze the error in the semi-discrete approximation (10.2). For simplicity, we assume (i) $a(\cdot,\cdot;t)$ is independent of t and (ii) $u^0 \in \mathcal{V}_h$.

Proposition 10.4. Suppose in the heat equation (10.1) the bilinear form $a(\cdot,\cdot;t)$ is independent of t and the initial condition u^0 is in $\mathcal{V}_h$. Let $\Pi_{A,h} : \mathcal{V} \to \mathcal{V}_h$ be the projection operator with respect to the symmetric, coercive bilinear form $a(\cdot,\cdot)$; i.e., for $w \in \mathcal{V}$, $\Pi_{A,h} w \in \mathcal{V}_h$ satisfies $a(w - \Pi_{A,h} w, v) = 0$ $\forall v \in \mathcal{V}_h$. Then the error in the semi-discrete form of the heat equation 10.2 is bounded by

$$\|u(t) - u_h(t)\|_{L^2(\Omega)} \leq \|(u - \Pi_{A,h} u)(t)\|_{L^2(\Omega)} + \int_{\tau=0}^{t} \exp(-\alpha(t-\tau)) \left\| \left. \frac{\partial (u - \Pi_{A,h} u)}{\partial t} \right|_\tau \right\|_{L^2(\Omega)} d\tau.$$

Proof. We first introduce $e_1 \equiv \Pi_{A,h} u - u_h$ and $e_2 \equiv u - \Pi_{A,h} u$ such that $u - u_h = e_1 + e_2$. We next note that, $\forall v \in \mathcal{V}_h$,

$$m\left(\left. \frac{\partial e_1}{\partial t} \right|_t, v \right) + a(e_1(t), v) = m\left(\left. \frac{\partial \Pi_{A,h} u}{\partial t} \right|_t, v \right) + \underbrace{\underbrace{a(\Pi_{A,h} u(t), v)}_{a(u,v)} \underbrace{- m\left(\left. \frac{\partial u_h}{\partial t} \right|_t, v \right) - a(u_h(t), v)}_{-\ell(v)}}_{-m(\frac{\partial u}{\partial t}, v)}$$

$$= -m\left(\left. \frac{\partial e_2}{\partial t} \right|_t, v \right).$$

We then note $e_1(t) \in \mathcal{V}_h$, $\forall t \in \mathcal{I}$, and choose $v = e_1(t)$ to obtain

$$\frac{1}{2} \frac{d}{dt} \|e_1(t)\|^2_{L^2(\Omega)} + a(e_1(t), e_1(t)) = - \left(\left. \frac{\partial e_2}{\partial t} \right|_t, e_1(t) \right)_{L^2(\Omega)}.$$

We then appeal to the coercivity of the bilinear form to obtain

$$\frac{1}{2} \frac{d}{dt} \|e_1(t)\|^2_{L^2(\Omega)} + \alpha \|e_1(t)\|^2_{H^1(\Omega)} \leq \left| \left(\left. \frac{\partial e_2}{\partial t} \right|_t, e_1(t) \right)_{L^2(\Omega)} \right|.$$

We then invoke (i) $\frac{1}{2}\|e_1(t)\|^2_{L^2(\Omega)} = \|e_1(t)\|_{L^2(\Omega)} \frac{d}{dt} \|e_1(t)\|^2_{L^2(\Omega)}$ on the first term, (ii) $\|e_1(t)\|_{H^1(\Omega)} \geq \|e_1(t)\|_{L^2(\Omega)}$ on the second term, (iii) $\left| \left(\left. \frac{\partial e_2}{\partial t} \right|_t, e_1(t) \right)_{L^2(\Omega)} \right| \leq \| \frac{\partial e_2}{\partial t} \|_{L^2(\Omega)} \|e_1(t)\|_{L^2(\Omega)}$ by Cauchy-Schwarz on the third term, and (iv) divide through by $\|e_1(t)\|_{L^2(\Omega)}$ to obtain

$$\frac{d}{dt} \|e_1(t)\|_{L^2(\Omega)} + \alpha \|e_1(t)\|_{L^2(\Omega)} \leq \left\| \left. \frac{\partial e_2}{\partial t} \right|_t \right\|_{L^2(\Omega)}$$

We integrate the ODE from $t = 0$ to t to obtain

$$\|e_1(t)\| \leq \exp(-\alpha t) \|e_1(t=0)\|_{L^2(\Omega)} + \int_{\tau=0}^{t} \exp(-\alpha(t-\tau)) \left\| \left. \frac{\partial e_2}{\partial t} \right|_\tau \right\|_{L^2(\Omega)} d\tau$$

For $u^0 \in \mathcal{V}_h$, $u_h(t=0) = \Pi_{A,h} u(t=0)$ and hence $e_1(t=0) = 0$. Hence

$$\begin{aligned}\|u(t) - u_h(t)\|_{L^2(\Omega)} &\leq \|e_1(t)\|_{L^2(\Omega)} + \|e_2(t)\|_{L^2(\Omega)} \\ &\leq \int_{\tau=0}^{t} \exp(-\alpha(t-\tau)) \left\| \left. \frac{\partial e_2}{\partial t} \right|_\tau \right\|_{L^2(\Omega)} d\tau + \|e_2(t)\|_{L^2(\Omega)},\end{aligned}$$

which is the desired relationship. □

We make a few remarks about Proposition 10.4. First, the error in the semi-discrete solution u_h at time t depends on two sources: (i) the projection error at time t; (ii) the projection errors committed in all previous times, as expressed as a time integral. Second, while the error at time t formally depends on all projection errors over $(0,t)$, the influence of the projection error at $t' < t$ on the solution $u_h(t)$ decays exponentially in time. Third, the particular rate of this decay depends on the coercivity constant α; the larger the coercivity constant, quickly the influence decays. Fourth, one limiting case of this observation is if the solution reaches a steady state as $t \to \infty$; in steady state, the error in the solution $u_h(t)$ depends only on the spatial projection error at the steady state.

We may also construct a particular bound for the piecewise polynomial space $\mathcal{V}_h = \{v \in \mathcal{V} \mid v|_K \in \mathbb{P}^p(K), \forall K \in \mathcal{T}_h\}$. We observe that if $u \in C^1(\bar{\mathcal{I}}; H^{s+1}(\Omega))$, then

$$\begin{aligned}\|(u - \Pi_{A,h} u)(t)\|_{L^2(\Omega)} &\leq C h^{r+1} |u(t)|_{H^{r+1}(\Omega)} \\ \|\frac{\partial (u - \Pi_{A,h} u)}{\partial t}\|_{L^2(\Omega)} &= \|\frac{\partial u}{\partial t} - \Pi_{A,h} \frac{\partial u}{\partial t}\|_{L^2(\Omega)} \leq C h^{r+1} |\frac{\partial u}{\partial t}|_{H^{r+1}(\Omega)}\end{aligned}$$

for $r = \min\{s, p\}$. It follows that, if $u \in C^1(\bar{\mathcal{I}}; H^{s+1}(\Omega))$, then

$$\|u(t) - u_h(t)\|_{L^2(\Omega)} \leq C h^{r+1}$$

for some constant C independent of h. We observe that if the weak solution is sufficiently regular (i.e,. $u \in C^1(\bar{\mathcal{I}}; H^{p+1}(\Omega))$), then the error in the semi-discrete solution u_h converges as h^{p+1} in the $L^2(\Omega)$ norm.

10.6 Full discrete formulation

We now apply a time integration scheme to the time derivative to obtain a full discrete form. We first introduce time steps $0 = t^0 < t^1 < \cdots < t^K = T$; the time steps need not be equispaced. We then seek a sequence of solutions $\hat{\hat{u}}^k$, $k = 0, \ldots, K$, such that $\hat{\hat{u}}^k \approx \hat{u}(t^k)$. For instance, we may apply a family of two-step integration schemes parameterized by $\theta \in [0,1]$ to the (algebraic) semi-linear form (10.3) to obtain

$$\begin{aligned}\frac{1}{\Delta t^k} \hat{M}_h (\hat{\hat{u}}_h^k - \hat{\hat{u}}_h^{k-1}) + \theta \hat{A}_h \hat{\hat{u}}_h^k + (1-\theta) \hat{A}_h \hat{\hat{u}}_h^{k-1} = \theta \hat{\hat{f}}_h^k + (1-\theta) \hat{\hat{f}}_h^{k-1} \quad \text{in } \mathbb{R}^n, \ k = 1, \ldots, K, \\ \hat{\hat{u}}_h^{k=0} = \hat{u}^0 \quad \text{in } \mathbb{R}^n,\end{aligned}$$

where $\Delta t^k \equiv t^k - t^{k-1}$. The choices of $\theta = 0$, $1/2$, and 1 yield the forward-Euler, Crank-Nicolson, and backward-Euler schemes, respectively.

As the focus of this course is on finite element methods and not ODE integration techniques, we will not discuss time integration in great depth; we simply make few remarks. First, there are

many other time-marching scheme that can be used; we refer to the AER336 course notes for some examples. Second, as the ODEs arising from the finite element discretization of parabolic equations are stiff — the condition number of $\hat{\hat{A}}_h$ scales as h^{-2} — implicit, and unconditionally stable, schemes are often used; the backward Euler and Crank-Nicolson schemes are the two classical choices. Third, due to the presence of the non-diagonal mass matrix $\hat{\hat{M}}_h$, a linear system must be solved even if an explicit time-integration scheme is used. Fourth, to circumvent this required linear solve, the mass matrix is often approximated by a diagonal matrix using a technique called "mass-lumping" *if* an explicit time marching scheme is used.

10.7 Full discrete formulation: error analysis

We now wish to assess the error in the full discrete solution $u^k_{h,\Delta t} \equiv \sum_{j=1}^n \hat{u}^k_{h,j}\phi_j$, $k = 1, \ldots, K$. To this end, we combine the error bound for the semi-discrete solution and for time-marching schemes to obtain, for spatially and temporally smooth solutions,

$$\|u(t^k) - u^k_{h,\Delta t}\|_{L^2(\Omega)} \leq \|u(t^k) - u_h(t^k)\|_{L^2(\Omega)} + \|u_h(t^k) - u^k_{h,\Delta t}\|_{L^2(\Omega)} \leq Ch^{p+1} + C'\Delta t^q,$$

where q is the order of accuracy of the time integration scheme (e.g., $q = 1$ for backward Euler). To control the error in $u_{h,\Delta t}$, we must control the spatial and temporal error by choosing sufficiently small h and Δt, respectively. For non-smooth solutions, the spatial and temporal convergence rates would be limited by the regularity of the solution.

10.8 Summary

We summarize key points of this lecture:

1. A weak formulation of parabolic equation is characterized by a mass bilinear form $m(\cdot,\cdot)$, a time-dependent bilinear form $a(\cdot,\cdot;t)$, and a time-dependent linear form $\ell(\cdot;t)$.

2. The energy $\|u(t)\|^2_{L^2(\Omega)}$ at time t is bounded by the energy at the initial time $\|u^0\|^2_{L^2(\Omega)}$ and the time-integrated data (source term) $\int_{\tau=0}^t \|\ell(\cdot;\tau)\|^2_{\mathcal{V}'} d\tau$.

3. A semi-discrete formulation is obtained by discretizing the spatial operator of the parabolic equation. If the solution is sufficiently smooth (i.e., $u \in C^1(\bar{\mathcal{I}}; H^{p+1}(\Omega))$) and a $\mathbb{P}^p$ finite element method is used, then the error in the semi-discrete formulation converges as Ch^{p+1} in $L^2(\Omega)$ norm.

4. A full discrete formulation is obtained by applying a time-marching scheme to a semi-discrete formulation. If the solution is sufficiently smooth (i.e., $u \in C^{q+1}(\bar{\mathcal{I}}; H^{p+1}(\Omega))$) and the formulation is based on a $\mathbb{P}^p$ finite element and a q-th order time integration scheme, then the error converges as $Ch^{p+1} + C'\Delta t^q$ in $L^2(\Omega)$ norm.

Lecture 11

Wave equation

11.1 Motivation

In this lecture we consider a variational formulation and the associated finite element approximation of the wave equation, which is a prototypical second-order hyperbolic equation that models the propagation of waves through a medium. Second-order hyperbolic equations are relevant in many engineering applications. In acoustics, the acoustic wave equation models the propagation of pressure waves. In structural dynamics, the elastodynamics equations model the dynamics of elastic structures. In electromagnetics, the Maxwell's equations model the propagation of electric and magnetic waves. In this lecture we focus on the model equation — the wave equation — to introduce numerical approximation of second-order hyperbolic equations.

11.2 Model problem: the wave equation

We first introduce our model second-order hyperbolic equation: the wave equation. To this end, we introduce a Lipschitz spatial domain $\Omega \in \mathbb{R}^d$ and a time interval $\mathcal{I} \equiv (0, T]$. We partition the domain boundary $\partial\Omega$ into the Dirichlet boundary Γ_D and the Neumann boundary Γ_N. We then introduce a wave equation

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} - \nabla^2 u &= f \quad \text{in } \Omega \times \mathcal{I}, \\ u &= 0 \quad \text{on } \Gamma_D \times \mathcal{I}, \\ \frac{\partial u}{\partial n} &= 0 \quad \text{on } \Gamma_N \times \mathcal{I}, \\ u|_{t=0} &= u^0 \quad \text{on } \Omega, \\ \left.\frac{\partial u}{\partial t}\right|_{t=0} &= u^1 \quad \text{on } \Omega. \end{aligned} \tag{11.1}$$

Note that because the equation is second-order in time, we require initial conditions on both the value and time-derivative. For simplicity we consider the problem with constant coefficients and homogeneous boundary conditions.

11.3 Weak formulation

We first introduce the space $\mathcal{V} \equiv H_0^1(\Omega)$, which is appropriate for our problem (11.1) with homogeneous Dirichlet data. We then multiply the strong form by a test function and integrate by parts to obtain a weak form of the wave equation: find $u(t) \in \mathcal{V}$, $t \in \mathcal{I}$, such that

$$\begin{aligned} m(\left.\frac{\partial^2 u}{\partial t^2}\right|_t, v) + a(u(t), v) &= \ell(v) \quad \forall v \in \mathcal{V}, \forall t \in \mathcal{I}, \\ m(u(t=0), v) &= m(u^0, v) \quad \forall v \in \mathcal{V}, \\ m(\left.\frac{\partial u}{\partial t}\right|_t, v) &= m(u^1, v) \quad \forall v \in \mathcal{V}, \end{aligned} \tag{11.2}$$

where the bilinear forms are given by

$$\begin{aligned} m(w, v) &\equiv \int_\Omega v w dx \quad \forall v, w \in \mathcal{V}, \\ a(w, v) &\equiv \int_\Omega \nabla v \cdot \nabla w dx \quad \forall v, w \in \mathcal{V}, \\ \ell(v) &\equiv \int_\Omega v f dx \quad \forall v \in \mathcal{V}. \end{aligned}$$

The solution to the wave equation (11.2) possesses the following energy conservation property.

Proposition 11.1 (energy conservation)**.** If $\ell(v) = 0 \ \forall v \in \mathcal{V}$, then the solution to the wave equation (11.2) conserves the total energy in the sense that

$$\frac{d}{dt}\Big(\underbrace{\frac{1}{2} m(\frac{\partial u}{\partial t}, \frac{\partial u}{\partial t})}_{\text{kinetic}} + \underbrace{\frac{1}{2} a(u, u)}_{\text{potential}} \Big) = 0.$$

Proof. We set $v = \frac{\partial u}{\partial t}$ in (11.2) and appeal to the symmetry of the forms $m(\cdot, \cdot)$ and $a(\cdot, \cdot)$ to obtain

$$0 = m(\frac{\partial^2 u}{\partial t^2}, \frac{\partial u}{\partial t}) + a(u, \frac{\partial u}{\partial t}) = \frac{d}{dt}\Big(\frac{1}{2} m(\frac{\partial u}{\partial t}, \frac{\partial u}{\partial t}) + \frac{1}{2} a(u, u) \Big),$$

which is the desired result. □

11.4 Semi-discrete formulation

We now introduce the semi-discrete formulation based on a finite-element spatial approximation. To this end, we introduce a finite element space

$$\mathcal{V}_h = \{ v \in \mathcal{V} \mid v|_K \in \mathbb{P}^p(K), \ \forall K \in \mathcal{T}_h \}.$$

Our semi-discrete formulation is as follows: find $u_h(t) \in \mathcal{V}_h$, $t \in \mathcal{I}$, such that

$$\begin{aligned} m\left(\left.\frac{\partial^2 u_h}{\partial t^2}\right|_t, v \right) + a(u_h(t), v) &= \ell(v) \quad \forall v \in \mathcal{V}_h, \\ m(u_h(t=0), v) &= m(u^0, v) \quad \forall v \in \mathcal{V}_h, \\ m\left(\left.\frac{\partial u_h}{\partial t}\right|_t, v \right) &= m(u^1, v) \quad \forall v \in \mathcal{V}_h. \end{aligned} \tag{11.3}$$

The solution to this semi-discrete equation also conserves the total energy; the proof follows that in Proposition 11.1.

To facilitate the presentation of the (fully) discrete equation, we now introduce the algebraic form of (11.3). We first introduce a basis $\{\phi_i\}_{i=1}^n$ of the space $\mathcal{V}_h$. We then represent the solution $u_h(x,t) = \sum_{j=1}^n \hat{u}_{h,j}(t)\phi_j(x)$, where $\hat{u}_h \in \mathbb{R}^n$ is the finite element vector. We now rewrite (11.3) as follows: find $\hat{u}_h(t) \in \mathbb{R}^n$, $t \in \mathcal{I}$, such that

$$m(\sum_{j=1}^n \left.\frac{d^2\hat{u}_{h,j}}{dt^2}\right|_t \phi_j, \phi_i) + a(\sum_{j=1}^n \hat{u}_{h,j}(t)\phi_j, \phi_i) = \ell(\phi_i) \quad \forall i = 1,\dots,n,\ t \in \mathcal{I}.$$

We then appeal to the bilinearity of $m(\cdot,\cdot)$ and $a(\cdot,\cdot)$ to obtain the following system of ODEs: find $\hat{u}_h(t) \in \mathbb{R}^n$, $t \in \mathcal{I}$, such that

$$\begin{aligned} \hat{M}_h \left.\frac{d^2\hat{u}_h}{dt^2}\right|_t + \hat{A}_h\hat{u}_h(t) &= \hat{f}_h \quad \text{in } \mathbb{R}^n,\ \forall t \in \mathcal{I}, \\ \hat{u}_h(t=0) &= \hat{u}^0 \quad \text{in } \mathbb{R}^n, \\ \left.\frac{d\hat{u}_h}{dt}\right|_{t=0} &= \hat{u}^1 \quad \text{in } \mathbb{R}^n, \end{aligned} \tag{11.4}$$

where the mass matrix $\hat{M}_h \in \mathbb{R}^{n\times n}$, stiffness matrix $\hat{A}_h \in \mathbb{R}^{n\times n}$, and load vector $\hat{f}_h \in \mathbb{R}^n$ are given by

$$\begin{aligned} \hat{M}_{h,ij} &= m(\phi_j,\phi_i), \quad i,j = 1,\dots,n, \\ \hat{A}_{h,ij} &= a(\phi_j,\phi_i), \quad i,j = 1,\dots,n, \\ \hat{f}_{h,i} &= \ell(\phi_i), \quad i = 1,\dots,n, \end{aligned}$$

and the initial condition vectors $\hat{u}^0 \in \mathbb{R}^n$ and $\hat{u}^1 \in \mathbb{R}^n$ satisfy

$$\begin{aligned} \hat{M}_h\hat{u}^0 &= \hat{h}^0 \quad \text{in } \mathbb{R}^n \\ \hat{M}_h\hat{u}^1 &= \hat{h}^1 \quad \text{in } \mathbb{R}^n \end{aligned}$$

for $\hat{h}_i^0 = (u^0,\phi_i)_{L^2(\Omega)}$ and $\hat{h}_i^1 = (u^1,\phi_i)_{L^2(\Omega)}$, $i = 1,\dots,n$.

11.5 First-order formulation and full discrete form

To obtain a full discrete form using a standard ODE time-integration scheme designed for first-order ODEs, we recast the second-order ODE in a first-order form. To this end, we introduce an auxiliary variable $\hat{v}_h \equiv \frac{\partial\hat{u}_h}{\partial t}$. The second-order ODE (11.4) is now recast as a system of (systems of) first-order ODEs:

$$\begin{aligned} \hat{M}_h \left.\frac{d\hat{u}_h}{dt}\right|_t - \hat{M}_h\hat{v}_h(t) &= 0 \quad \text{in } \mathbb{R}^n,\ \forall t \in \mathcal{I}, \\ \hat{M}_h \left.\frac{d\hat{v}_h}{dt}\right|_t + \hat{A}_h\hat{u}_h(t) &= \hat{f}_h \quad \text{in } \mathbb{R}^n,\ \forall t \in \mathcal{I}, \\ \hat{u}_h(t=0) &= \hat{u}^0 \quad \text{in } \mathbb{R}^n, \\ \hat{v}_h(t=0) &= \hat{u}^1 \quad \text{in } \mathbb{R}^n. \end{aligned} \tag{11.5}$$

The system can be written more compactly written using a block matrix

$$\begin{pmatrix} \hat{M}_h & 0 \\ 0 & \hat{M}_h \end{pmatrix} \frac{d}{dt} \begin{pmatrix} \hat{u}_h \\ \hat{v}_h \end{pmatrix} + \begin{pmatrix} 0 & -\hat{M}_h \\ \hat{A}_h & 0 \end{pmatrix} \begin{pmatrix} \hat{u}_h \\ \hat{v}_h \end{pmatrix} = \begin{pmatrix} 0 \\ \hat{f}_h \end{pmatrix} \quad \text{in } \mathbb{R}^{2n}.$$

This first-order system can be solved using any time-integration scheme.

We now briefly discuss the choice of a time integration scheme. (Similar to the lecture on parabolic equations, we refer to the AER336 course notes for more detailed treatment of time integration schemes.) One classical choice of time integrator for the wave equation is the Crank-Nicolson method, which yields an algebraic system of the form

$$\begin{pmatrix} \hat{M}_h & -\frac{\Delta t}{2}\hat{M}_h \\ \frac{\Delta t}{2}\hat{A}_h & \hat{M}_h \end{pmatrix} \begin{pmatrix} \hat{u}_h^{k+1} \\ \hat{v}_h^{k+1} \end{pmatrix} = \begin{pmatrix} \hat{M}_h & \frac{\Delta t}{2}\hat{M}_h \\ -\frac{\Delta t}{2}\hat{A}_h & \hat{M}_h \end{pmatrix} \begin{pmatrix} \hat{u}_h^{k} \\ \hat{v}_h^{k} \end{pmatrix} + \begin{pmatrix} 0 \\ \Delta t \hat{f}_h \end{pmatrix} \quad \text{in } \mathbb{R}^{2n}.$$

The Crank-Nicolson method is well-suited for the wave equation, as the full discrete system, just like the continuous counterpart, conserves the total energy in the system. The Crank-Nicolson method applied to the wave equation conserves energy because (i) the eigenvalues of the matrix $\begin{pmatrix} 0 & -\hat{M}_h \\ \hat{A}_h & 0 \end{pmatrix}$ are purely imaginary and (ii) the stability boundary for the Crank-Nicolson method is along the imaginary axis.

11.6 Error analysis

For smooth problems, it can be shown that the $\mathbb{P}^p$ finite element spatial discretization and the Crank-Nicolson time integration yields

$$\|u(t) - u_{h,\Delta t}(t)\|_{L^2(\Omega)} \leq C_1 h^{p+1} + C_2 \Delta t^2,$$
$$\|u(t) - u_{h,\Delta t}(t)\|_{H^1(\Omega)} \leq C_1 h^{p} + C_2 \Delta t^2.$$

In other words, the scheme is $p+1$-st order in space in $L^2(\Omega)$ and second-order in time. In general, if the initial condition and/or the data f is not smooth, then the scheme still converges but at a lower convergence rate.

11.7 Generalization to other second-order hyperbolic equations

As noted in the introduction, the wave equation is a model equation for second-order hyperbolic equations. In fact, because of the abstraction provided by the weak formulation (11.2), we can simply redefine (i) the function space and (ii) bilinear forms to obtain various equations. We here provide a few examples.

Acoustic wave equation. The solution field for the acoustic wave equation is the time-dependent pressure (perturbation) field. As the pressure field is a scalar field we consider $\mathcal{V}$ such that $H_0^1(\Omega) \subset \mathcal{V} \subset H^1(\Omega)$. The bilinear forms are given by

$$m(w,v) \equiv \int_\Omega \frac{1}{c^2} wv dx \quad \forall w, v \in \mathcal{V},$$
$$a(w,v) \equiv \int_\Omega \nabla v \cdot \nabla w dx \quad \forall w, v \in \mathcal{V},$$

where $c > 0$ is the speed of sound.

Elastodynamics. The solution field for the elastodynamics equation is the time-dependent displacement field. As we have seen in the lecture on linear elasticity, we consider a vector-valued field $\mathcal{V}$ such that $H_0^1(\Omega)^d \subset \mathcal{V} \subset H^1(\Omega)^d$. The bilinear forms are given by

$$m(w, v) \equiv \int_\Omega \rho v \cdot w dx \quad \forall w, v \in \mathcal{V},$$
$$a(w, v) \equiv \int_\Omega (2\mu\epsilon(v) : \epsilon(w) + \lambda \mathrm{tr}(\epsilon(v))\mathrm{tr}(\epsilon(w)))dx \quad \forall w, v \in \mathcal{V},$$

where ρ is the density (field), $\epsilon(v) = \frac{1}{2}(\nabla v + \nabla v^T)$ is the strain tensor, and $\lambda \in L^\infty(\Omega)$ and $\mu \in L^\infty(\Omega)$ are the first and second Lamé parameters, respectively.

We make two remarks. First, the weak formulation must be completed by incorporating the particular boundary conditions associated with the physical problem; for instance, in acoustics the boundary condition depends . Second, because $m(\cdot, \cdot)$ is symmetric and positive and $a(\cdot, \cdot)$ is symmetric and coercive, both the acoustic wave equation and the elastodynamics equations share much of mathematical properties of "the" wave equation studied in this lecture; for instance, the total energy is conserved for both systems following Proposition 11.1.

11.8 Summary

We summarize key points of this lecture:

1. A weak formulation of the second-order hyperbolic equation is characterized by a mass bilinear form $m(\cdot, \cdot)$ and a spatial bilinear form $a(\cdot, \cdot)$.

2. The solution to the wave equation preserves the total energy, which is the sum of the kinetic and potential energies.

3. A classical approach to discretize the wave equation is to apply a $\mathbb{P}^p$ finite element approximation in space, rewrite the second-order ODE as a system of first-order ODEs, and then apply the Crank-Nicolson time integration. The resulting approximation preserves the total energy and, assuming the solution is smooth, is $p+1$-st order in space in $L^2(\Omega)$ and second-order in time.

4. Equations in continuum mechanics that have the same mathematical structure as the wave equation include the acoustic wave equation and the elastodynamics equations.

Lecture 12

Discontinuous Galerkin methods

12.1 Motivation

In the previous lecture, we observed that the standard Galerkin method is not well-suited for hyperbolic or advection-dominated problems and devised a stabilized method: the Galerkin least-squares method. In this lecture we consider an alternative formulation: the discontinuous Galerkin (DG) method.

12.2 Problem statement

We first introduce a Lipschitz domain $\Omega \subset \mathbb{R}^d$, a time interval $\mathcal{I} \equiv (0, T]$, an advection field $b \in L^\infty(\Omega)^d$, and the associated inflow and outflow boundaries:

$$\begin{aligned} \Gamma_{\text{in}} &\equiv \{x \in \partial\Omega \mid n(x) \cdot b(x) \leq 0\}, \\ \Gamma_{\text{out}} &\equiv \partial\Omega \setminus \Gamma_{\text{in}}. \end{aligned}$$

The strong form of a general hyperbolic equation is given by

$$\begin{aligned} \frac{\partial u}{\partial t} + \nabla \cdot f(u) &= 0 \quad \text{in } \Omega \times \mathcal{I}, \\ u &= u^b \quad \text{on } \Gamma_{\text{in}} \times \mathcal{I}, \\ u(t=0) &= u^0 \quad \text{in } \Omega, \end{aligned} \tag{12.1}$$

where $f(u)$ is the flux function, which is $f(u) = bu$ for advection equation. Note that the boundary condition is imposed only on the inflow boundary.

12.3 Discontinuous Galerkin method

To introduce the DG method, we first introduce a triangulation $\mathcal{T}_h = \{K\}$ of domain Ω such that $\sum_{K \in \mathcal{T}_h} \bar{K} = \bar{\Omega}$ and $K \cap K' = \emptyset \; \forall K, K' \in \mathcal{T}_h$. We also denote the set of all facets of the triangulation by $\Sigma_h = \{\sigma\}$. We further decompose the facet set Σ_h into the boundary facet set $\Sigma_h^{\text{b}} = \Sigma_h \cap \partial\Omega$ and the interior facet set $\Sigma_h^{\text{i}} = \Sigma_h \setminus \Sigma_h^{\text{b}}$. On each interior facet, we arbitrarily assign one of the elements abutting the facet as the "+" element and the other element as the "−" element. (The

DG scheme will be independent of this arbitrary assignment.) We then introduce an associated space of discontinuous piecewise polynomials

$$\mathcal{V}_h \equiv \{v \in L^2(\Omega) \mid v|_K \in \mathbb{P}^p(K), \forall K \in \mathcal{T}_h\}.$$

We only require the functions to be in $L^2(\Omega)$ and not $H^1(\Omega)$ and hence the space $\mathcal{V}_h$ contains discontinuous functions. As the functions in the space are discontinuous, the functions are "double-valued" on the interior facets; to circumvent the ambiguity, we denote the function $w \in \mathcal{V}_h$ evaluated on the $+$ and $-$ elements by w^+ and w^-, respectively.

To derive a DG method we first multiply the strong form (12.1) by $v_h \in \mathcal{V}_h$ and integrate by parts on each $K \in \mathcal{T}_h$:

$$\int_K v_h \frac{\partial u_h}{\partial t} dx + \int_K v_h \nabla \cdot f(u_h) dx = \int_K v_h \frac{\partial u_h}{\partial t} dx - \int_K \nabla v_h \cdot f(u_h) dx + \int_{\partial K} v_h^+ n^+ \cdot f(u_h^+) ds = 0,$$

where, for notational simplicity, we have assumed that the element K is on the "+" side of each of the abutting facets. We then replace the flux on the interface with a *numerical flux*. For an interior facet $\sigma \in \Sigma_h^{\mathrm{i}}$, the numerical flux depends on the state on both sides of the facet and is of the form

$$\hat{f}(w^+, w^-; n^+) = \frac{1}{2} n^+ \cdot (f(w^+) + f(w^-)) + \frac{1}{2} c(w^+, w^-; n^+)(w^+ - w^-);$$

here the function $c(w^+, w^-; n^+)$ is chosen such that the numerical flux is

i. *consistent*: $f(w, w; n^+) = n^+ \cdot f(w)$,

ii. *conservative*: $f(w^+, w^-; n^+) = -f(w^-, w^+; n^-)$,

iii. *dissipative*: $[w]_-^+ \hat{f}(w^+, w^-; n^+) \geq 0$,

where $[w]_-^+ = w^+ - w^-$. For instance, for the advection equation with $f(u) = bu$, a common choice is $c(w^+, w^-; n^+) = |b \cdot n^+|$, which yields the "upwinding" flux:

$$\hat{f}(w^+, w^-; n^+) = \begin{cases} n^+ \cdot f(w^+), & b \cdot n^+ \geq 0, \\ n^+ \cdot f(w^-), & b \cdot n^+ < 0; \end{cases}$$

in words, our numerical flux is based on the function value upwind of the facet. For a boundary facet $\sigma \in \Sigma_h^{\mathrm{b}}$, the numerical flux is given by

$$\hat{f}^{\mathrm{b}}(w^+, u^{\mathrm{b}}; n^+) = \begin{cases} n^+ \cdot f(u^b) & \text{on } \Gamma_{\mathrm{in}} \\ n^+ \cdot f(w^+) & \text{on } \Gamma_{\mathrm{out}} \end{cases};$$

the boundary flux again results from upwinding. With these choices of the numerical fluxes, the element-wise DG residual statement becomes

$$\int_K v_h \frac{\partial u_h}{\partial t} dx - \int_K \nabla v_h \cdot f(u_h) dx + \int_{\partial K \cap \Sigma_h^{\mathrm{i}}} v_h^+ \hat{f}(u_h^+, u_h^-; n^+) ds + \int_{\partial K \cap \Sigma_h^{\mathrm{b}}} v_h^+ \hat{f}^{\mathrm{b}}(u_h^+, u^{\mathrm{b}}; n^+) ds = 0.$$

We now sum over all $K \in \mathcal{T}_h$ to obtain the global DG residual statement: find $u_h(t) \in \mathcal{V}_h$, $t \in \mathcal{I}$, such that

$$\sum_{K\in\mathcal{T}_h}\left(\int_K v_h\frac{\partial u_h}{\partial t}dx - \int_K \nabla v_h \cdot f(u_h)dx\right) + \sum_{\sigma\in\Sigma_h^{\mathrm{i}}}\int_\sigma [v_h]_-^+ \hat{f}(u_h^+, u_h^-; n^+)ds$$
$$+ \sum_{\sigma\in\Sigma_h^{\mathrm{b}}}\int_\sigma v_h^+ \hat{f}^{\mathrm{b}}(u_h^+, u^{\mathrm{b}}; n^+)ds = 0 \quad \forall v_h \in \mathcal{V}_h,$$

where we have appealed to the conservativity of the numerical flux to combine the interior facet terms for the + and − elements. The statement may be more compactly stated as follows: find $u_h(t) \in \mathcal{V}$, $t \in \mathcal{I}$, such that

$$\int_\Omega v_h\frac{\partial u_h}{\partial t}dx + r_h(u_h, v_h) = 0 \quad \forall v_h \in \mathcal{V}_h,$$

where the spatial residual form is given by

$$r_h(w, v) \equiv -\int_{\mathcal{T}_h} \nabla v \cdot f(w)dx + \int_{\Sigma_h^{\mathrm{i}}} [v]_-^+ \hat{f}(w^+, w^-; n^+)ds + \int_{\Sigma_h^{\mathrm{b}}} v^+ \hat{f}^{\mathrm{b}}(w^+, u^{\mathrm{b}}; n^+)ds.$$

12.4 Energy-stability analysis for linear equations

We now analyze the energy stability of a linear advection equation given by the flux $f(u) \equiv bu$:

$$\frac{\partial u}{\partial t} + \nabla \cdot (bu) = 0 \quad \text{in } \Omega \times \mathcal{I}. \tag{12.2}$$

For simplicity, we assume that the advection field b is divergence-free: $\nabla \cdot b = 0$.

We first analyze the energy stability of the exact problem (12.2).

Proposition 12.1. The advection equation is energy stable (modulo boundary condition) with the following energy balance:

$$\frac{1}{2}\int_\Omega u(t=T)^2dx + \frac{1}{2}\int_{\mathcal{I}}\int_{\Gamma_{\mathrm{out}}} |b\cdot n|u^2dsdt = \frac{1}{2}\int_\Omega u(t=0)^2dx + \frac{1}{2}\int_{\mathcal{I}}\int_{\Gamma_{\mathrm{in}}} |b\cdot n|u^2dsdt,$$

where $\frac{1}{2}\|u\|^2_{L^2(\Omega)}$ is the energy in the system. In other words,

$$(\text{energy at } t=T) + (\text{energy leaving } \Omega) = (\text{energy at } t=0) + (\text{energy entering } \Omega);$$

i.e., the energy is conserved in the advection equation (modulo the boundary conditions).

Proof. We multiply the advection equation (12.2) by the solution u and integrate in both space and time

$$\int_{\mathcal{I}}\int_\Omega u\frac{\partial u}{\partial t}dxdt + \int_{\mathcal{I}}\int_\Omega u\nabla\cdot(bu)dxdt = 0.$$

The time derivative term simplifies to

$$\begin{aligned}\int_{\mathcal{I}}\int_{\Omega} u\frac{\partial u}{\partial t}dxdt &= \int_{\mathcal{I}}\int_{\Omega}\frac{\partial}{\partial t}\left(\frac{1}{2}u^2\right)dxdt = \int_{\mathcal{I}}\frac{d}{dt}\left(\frac{1}{2}\int_{\Omega}u^2dx\right)dt \\ &= \frac{1}{2}\int_{\Omega}u(t=T)^2dx - \frac{1}{2}\int_{\Omega}u(t=0)^2dx\end{aligned}$$

To simplify the advection term, we first note that

$$\begin{aligned}\int_{\Omega}u\nabla\cdot(bu)dx &= \frac{1}{2}\int_{\Omega}ub\cdot\nabla udx - \frac{1}{2}\int_{\Omega}\nabla u\cdot budx + \frac{1}{2}\int_{\partial\Omega}(b\cdot n)u^2ds \\ &= \frac{1}{2}\int_{\partial\Omega}(b\cdot n)u^2ds = \frac{1}{2}\int_{\Gamma_{\text{out}}}|b\cdot n|u^2ds - \frac{1}{2}\int_{\Gamma_{\text{in}}}|b\cdot n|u^2ds,\end{aligned}$$

where the first equality uses $\nabla\cdot b = 0$. We then integrate the advection term over $\mathcal{I}$ to obtain the desired relationship. □

We now analyze the energy stability of the DG approximation. To this end, we first note that the flux is given by $f(u) = bu$ and then set the $c(w^+, w^-; n^+)$ term of the interior flux to $c = |b\cdot n|$. With this choice, we obtain the following energy balance.

Proposition 12.2. The DG approximation of the advection equation is energy stable (modulo boundary condition) with the following energy balance:

$$\begin{aligned}&\frac{1}{2}\int_{\Omega}u_h^2(t=T) + \frac{1}{2}\int_{\mathcal{I}}\int_{\Sigma_h^{\text{i}}}[u_h]_-^+|b\cdot n|[u_h]_-^+dsdt + \frac{1}{2}\int_{\mathcal{I}}\int_{\Sigma_h^{\text{b}}\cap\Gamma_{\text{out}}}u_h^+|b\cdot n|u_h^+ds \\ &\quad= \frac{1}{2}\int_{\Omega}u_h^2(t=0) + \frac{1}{2}\int_{\mathcal{I}}\int_{\Sigma_h^{\text{b}}\cap\Gamma_{\text{in}}}u_h^+|b\cdot n|u_h^{\text{b}}dsdt.\end{aligned}$$

In other words,

$$\frac{1}{2}\int_{\Omega}u_h^2(t=T) \le \frac{1}{2}\int_{\Omega}u_h^2(t=0) + \frac{1}{2}\int_{\mathcal{I}}\int_{\Sigma_h^{\text{b}}\cap\Gamma_{\text{in}}}u_h^+|b\cdot n|u_h^{\text{b}}dsdt.$$

Proof. We first observe that

$$\begin{aligned}
r(u_h,u_h) &= -\int_{\mathcal{T}_h} u_h\cdot f(u_h)dx + \int_{\Sigma_h^{\mathrm{i}}} [u_h]_-^+ \hat{f}(u_h^+,u_h^-;n^+)ds + \int_{\Sigma_h^{\mathrm{b}}} u_h^+\hat{f}^{\mathrm{b}}(u_h^+,u^{\mathrm{b}};n^+)ds \\
&= \sum_{K\in\mathcal{T}_h}\left(-\frac{1}{2}\int_K \nabla u_h\cdot f(u_h)dx + \frac{1}{2}\int_K u_h\nabla\cdot f(u_h)dx - \frac{1}{2}\int_{\partial K} u_h n\cdot f(u_h)ds\right) \\
&\quad + \int_{\Sigma_h^{\mathrm{i}}} [u_h]_-^+ \hat{f}(u_h^+,u_h^-;n^+)ds + \int_{\Sigma_h^{\mathrm{b}}} u_h^+\hat{f}^{\mathrm{b}}(u_h^+,u^{\mathrm{b}};n^+)ds \\
&= \int_{\mathcal{T}_h} \underbrace{(-\frac{1}{2}\nabla u_h\cdot f(u_h) + \frac{1}{2}u_h\nabla\cdot f(u_h))}_{\text{(I)}}\,dx \\
&\quad + \int_{\Sigma_h^{\mathrm{i}}} \underbrace{(-\frac{1}{2}u_h^+ n^+\cdot f(u_h^+) - \frac{1}{2}u_h^- n^-\cdot f(u_h^-) + [u_h]_-^+\hat{f}(u_h^+,u_h^-;n^+))}_{\text{(II)}}\,ds \\
&\quad + \int_{\Sigma_h^{\mathrm{b}}\cap\Gamma_{\mathrm{in}}} \underbrace{(-\frac{1}{2}u_h^+ n^+\cdot f(u_h^+) + u_h^+\hat{f}^{\mathrm{b}}(u_h^+,u^{\mathrm{b}};n^+))}_{\text{(III)}}\,ds \\
&\quad + \int_{\Sigma_h^{\mathrm{b}}\cap\Gamma_{\mathrm{out}}} \underbrace{(-\frac{1}{2}u_h^+ n^+\cdot f(u_h^+) + u_h^+\hat{f}^{\mathrm{b}}(u_h^+,u^{\mathrm{b}};n^+))}_{\text{(IV)}}\,ds.
\end{aligned}$$

We now analyze the terms (I), (II), (III), and (IV). We first observe that

$$\text{(I)} = -\frac{1}{2}\nabla u_h\cdot f(u_h) + \frac{1}{2}u_h\cdot\nabla\cdot f(u_h) = -\frac{1}{2}\nabla u_h\cdot bu_h + \frac{1}{2}u_h\cdot\nabla\cdot(bu_h) = 0$$

for a divergence-free advection field b. We next observe that the interface facet term simplifies as

$$\begin{aligned}
\text{(II)} &= -\frac{1}{2}u_h^+ n^+\cdot f(u_h^+) - \frac{1}{2}u_h^- n^-\cdot f(u_h^-) + [u_h]_-^+\hat{f}(u_h^+,u_h^-;n^+) \\
&= -\frac{1}{2}u_h^+ n^+\cdot(bu_h^+) - \frac{1}{2}u_h^- n^-\cdot(bu_h^-) + [u_h]_-^+(\frac{1}{2}n^+\cdot(bu_h^+ + bu_h^-) + \frac{1}{2}|b\cdot n|[u_h]_-^+) \\
&= \frac{1}{2}(-u_h^+ n^+\cdot(bu_h^+) - u_h^- n^-\cdot(bu_h^-) + u_h^+ n^+\cdot(bu_h^+ + bu_h^-) + u_h^- n^-\cdot(bu_h^+ + bu_h^-) + [u_h]_-^+|b\cdot n|[u_h]_-^+) \\
&= \frac{1}{2}[u_h]_-^+|b\cdot n|[u_h]_-^+.
\end{aligned}$$

We then observe that the inflow boundary facet term simplifies to

$$\text{(III)} = -\frac{1}{2}u_h^+ n^+\cdot f(u_h^+) + u_h^+\hat{f}^{\mathrm{b}}(u_h^+,u^{\mathrm{b}};n^+) = -\frac{1}{2}u_h^+ n^+\cdot(bu^{\mathrm{b}}) + u_h^+ n^+\cdot bu^{\mathrm{b}} = -\frac{1}{2}u_h^+|b\cdot n|u_h^{\mathrm{b}},$$

where the last equality follows from $b\cdot n^+ = -|b\cdot n|$ on Γ_{in}. Similarly, the outflow boundary facet term simplifies to

$$\text{(IV)} = -\frac{1}{2}u_h^+ n^+\cdot f(u_h^+) + u_h^+\hat{f}^{\mathrm{b}}(u_h^+,u^{\mathrm{b}};n^+) = -\frac{1}{2}u_h^+ n^+\cdot(bu_h^+) + u_h^+ n^+\cdot bu_h^+ = \frac{1}{2}u_h^+|b\cdot n|u_h^+,$$

where the last equality follows from $b \cdot n^+ = |b \cdot n|$ on Γ_{out}. It follows that

$$r(u_h, u_h) = \frac{1}{2}\int_{\Sigma_h^{\text{i}}} [u_h]_-^+ |b \cdot n| [u_h]_-^+ ds - \frac{1}{2}\int_{\Sigma_h^{\text{b}} \cap \Gamma_{\text{in}}} u_h^+ |b \cdot n| u_h^{\text{b}} ds + \frac{1}{2}\int_{\Sigma^{\text{b,out}} \cap \Gamma_{\text{out}}} u_h^+ |b \cdot n| u_h^+ ds.$$

The integration of the residual term over $\mathcal{I}$ yields

$$\begin{aligned}
&\frac{1}{2}\int_\Omega u_h^2(t=T) + \frac{1}{2}\int_{\mathcal{I}}\int_{\Sigma_h^{\text{i}}} [u_h]_-^+ |b \cdot n| [u_h]_-^+ ds dt + \frac{1}{2}\int_{\mathcal{I}}\int_{\Sigma_h^{\text{b}} \cap \Gamma_{\text{out}}} u_h^+ |b \cdot n| u_h^+ ds \\
&\quad = \frac{1}{2}\int_\Omega u_h^2(t=0) + \frac{1}{2}\int_{\mathcal{I}}\int_{\Sigma_h^{\text{b}} \cap \Gamma_{\text{in}}} u_h^+ |b \cdot n| u_h^{\text{b}} ds dt,
\end{aligned}$$

which is the desired result. □

We observe that in the DG formulation the jump term on the interior facet $\frac{1}{2}\int_{\mathcal{I}}\int_{\Sigma_h^{\text{i}}} [u_h]_-^+ |b \cdot n| [u_h]_-^+ ds dt$ provides additional dissipation relative to the original weak formulation. This dissipation ensures that the DG formulation is energy stable.

12.5 Observations

We make a few additional observations about the DG method:

- For $\mathbb{P}^{p=0}$, the DG method reduces to the finite volume method without reconstruction.
- The DG method is locally conservative:

  $$\frac{d}{dt}\int_K u_h dx + \int_{\partial\kappa} \hat{f}(u_h^+, u_h^-; n^+) ds = 0 \quad \forall K \in \mathcal{T}_h.$$

  The result is a direct consequence of the test space which includes element-wise constant functions.
- The mass matrix is block diagonal. The block-diagonal mass matrix enables efficient implementation of explicit time-marching schemes. The block-diagonal mass matrix is a direct consequence of the discontinuous approximation space.
- *A priori* error analysis: if $u \in H^{s+1}(\Omega)$, then

  $$\|u - u_h\|_{L^2(\Omega)} \leq C h^{r+1/2} |u|_{H^{r+1}(\Omega)},$$

  for $r \equiv \min\{s, p\}$. This formal convergence rate is suboptimal by the order of $1/2$; however, in practice, we almost always observe the optimal convergence rate of h^{p+1} for smooth problems.
- The DG method is energy stable for linear hyperbolic systems for an appropriate choice of the numerical fluxes (as analyzed in Section 12.4).
- The DG method is entropy stable for nonlinear hyperbolic systems for an appropriate choice of the numerical fluxes.

12.6 DG methods for elliptic equations (brief overview)

DG methods can treat also elliptic PDEs. To illustrate a formulation using a concrete example, we introduce a Poisson's problem in $\Omega \subset \mathbb{R}^d$:

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= u^b \quad \text{on } \Gamma_D, \\ \frac{\partial u}{\partial n} &= g \quad \text{on } \Gamma_N. \end{aligned}$$

There are several different DG discretizations for the elliptic equations. For simplicity, we consider the interior penalty (IP) method.

To introduce the IP-DG method, we first recall the DG approximation space

$$\mathcal{V}_h \equiv \{v \in L^2(\Omega) \mid v|_K \in \mathbb{P}^p(K), \forall K \in \mathcal{T}_h\}.$$

We note that, unlike the finite element space for the standard Galerkin discretization, the DG space does *not* incorporate the Dirichlet boundary condition. We then define for $v_h \in \mathcal{V}_h$ the facet jump operator

$$[\![v]\!] = \begin{cases} v^+ n^+ + v^- n^- & \text{on } \Gamma_h^i, \\ v^+ & \text{on } \Gamma_h^b \equiv \Gamma_D \cup \Gamma_N, \end{cases}$$

and the facet averaging operator

$$\{v\} = \begin{cases} \frac{1}{2}(v^+ - v^-) & \text{on } \Gamma_h^i, \\ v^+ & \text{on } \Gamma_h^b. \end{cases}$$

The IP-DG approximation of the Poisson's problem is as follows: find $u_h \in \mathcal{V}_h$ such that

$$a_h(u_h, v_h) = \ell_h(v_h) \quad \forall v_h \in \mathcal{V}_h,$$

where

$$\begin{aligned} a_h(w, v) &\equiv \int_\Omega \nabla v \cdot \nabla w dx - \int_{\Sigma_h^i \cup \Gamma_D} (\{\nabla v\} \cdot [\![w]\!] + [\![v]\!] \cdot \{\nabla w\}) ds + \int_{\Sigma_h^i \cup \Gamma_D} \vartheta [\![v]\!] \cdot [\![w]\!] ds, \\ \ell_h(v) &\equiv \int_\Omega v f dx + \int_{\Gamma_N} v g ds - \int_{\Gamma_D} (\nabla v^+ \cdot n) u^b ds + \int_{\Gamma_D} \vartheta v^+ u^b ds, \end{aligned}$$

and $\vartheta \in \mathbb{R}_{>0}$ is the stabilization parameter that scales as $\vartheta \sim 1/h$. The bilinear form comprises the standard Galerkin term on Ω and the DG-specific terms on $\Sigma_h^i \cup \Gamma_D$. Similarly, the linear form comprises the standard Galerkin terms on Ω and Γ_N and the DG-specific terms on Γ_D. We note that both the Dirichlet and Neumann boundary conditions are weakly enforced.

We conclude this short section with a few remarks:

- Let the DG-norm by given by

  $$\|v\|_{\mathrm{DG}}^2 \equiv \sum_{\kappa \in \mathcal{T}_h} \|\nabla v\|_{L^2(\kappa)}^2 + \int_{\Sigma_h^i \cup \Gamma_D} (\vartheta |[\![v]\!]|^2 + \vartheta^{-1} |\{\nabla v\}|^2) ds.$$

  The bilinear form $a_h(\cdot, \cdot)$ is coercive with respect to $\|\cdot\|_{\mathrm{DG}}$ in $\mathcal{V}_h$ assuming $\vartheta > C/h$, where the minimum value of C depends on the shape of the elements.

- The IP-DG method is (primal) consistent and adjoint consistent.
- The DG error converges at the optimal rate: if $u \in H^{s+1}(\Omega)$, then $\|u - u_h\|_{\mathrm{DG}} < Ch^r$ for $r = \min\{s, p\}$.

12.7 Summary

We summarize key points of this lecture:

1. The DG method provides energy/entropy stable approximations of hyperbolic equations.
2. The DG method seeks solution in discontinuous element-wise polynomial spaces.
3. The choice of the numerical flux — which should be consistent, conservative, and dissipative — plays a crucial role in ensuring the energy/entropy stability of the DG formulation. The dissipation is provided by the jump in the state across elements.
4. We refer to Section 12.5 for a bullet list of additional properties of the DG method for hyperbolic PDEs.
5. DG methods can also treat elliptic PDEs.

Lecture 13

Navier-Stokes equations

13.1 Motivation

In this lecture we consider a weak formulation and the associated finite element approximation of the incompressible Navier-Stokes equations. The Navier-Stokes equations play an important role in the design and analysis of fluids systems and multi-physics systems that involve fluids, such as fluid-thermal transport and fluid-structure interaction. The Navier-Stokes equations also allow us to exercise the formulation and implementation of finite element methods for nonlinear and non-coercive PDEs.

13.2 Strong and weak formulations

Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. We partition the boundary $\partial\Omega$ into an inflow boundary Γ_{in}, outflow boundary Γ_{out}, and (no-slip) wall boundary Γ_{wall}, such that $\overline{\partial\Omega} = \overline{\Gamma}_{\text{in}} \cup \overline{\Gamma}_{\text{out}} \cup \overline{\Gamma}_{\text{wall}}$ and $\Gamma_{\text{in}} \cap \Gamma_{\text{out}} = \Gamma_{\text{in}} \cap \Gamma_{\text{wall}} = \Gamma_{\text{out}} \cap \Gamma_{\text{wall}} = \emptyset$. The strong form of the Navier-Stokes equations for the (vector-valued) velocity u and (scalar-valued) pressure p is given by

$$\begin{aligned} \frac{\partial u}{\partial t} + (u \cdot \nabla)u - \nu\Delta u + \nabla p = 0 \quad &\text{in } \Omega, \\ \nabla \cdot u = 0 \quad &\text{in } \Omega, \end{aligned}$$

where the first (vector-valued) equation is associated with the conservation of momentum, the second (scalar-valued) equation is associated with the conservation of mass, and $\nu \in \mathbb{R}_{>0}$ is the kinematic viscosity. The first equation is often called *momentum equation*; the second equation is often called *continuity equation* or *divergence-free condition*. The equations are augmented by the following boundary conditions:

$$\begin{aligned} u = u_{\text{in}} \quad &\text{on } \Gamma_{\text{in}}, \\ \nu(n \cdot \nabla u) - pn = 0 \quad &\text{on } \Gamma_{\text{out}}, \\ u = 0 \quad &\text{on } \Gamma_{\text{wall}}. \end{aligned}$$

(We can readily consider other boundary conditions, but we focus on these three boundary conditions in this lecture (i) to simplify the presentation and (ii) as they are frequently encountered in practice.) The equations can also be written using index notation:

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} - \nu \frac{\partial^2 u_i}{\partial x_j \partial x_j} + \frac{\partial p}{\partial x_i} = 0 \quad \text{in } \Omega,$$
$$\frac{\partial u_j}{\partial x_j} = 0 \quad \text{in } \Omega,$$

with boundary conditions

$$u_i = u_{\text{in},i} \quad \text{on } \Gamma_{\text{in}},$$
$$\nu n_j \frac{\partial u_i}{\partial x_j} - p n_i = 0 \quad \text{on } \Gamma_{\text{out}},$$
$$u_i = 0 \quad \text{on } \Gamma_{\text{wall}};$$

here the equations are enforced for $i = 1, \ldots, d$, and the summation on repeated j indices are implied.

To obtain the weak form of the Navier-Stokes equations, we first introduce approximation spaces for the velocity and pressure,

$$\mathcal{V} \equiv \{v \in H^1(\Omega)^d \mid v|_{\Gamma_{\text{wall}} \cup \Gamma_{\text{in}}} = 0\},$$
$$\mathcal{Q} \equiv L^2(\Omega),$$

and the associated velocity space with nonhomogeneous boundary conditions

$$\mathcal{V}^E \equiv \{v \in H^1(\Omega)^d \mid v|_{\Gamma_{\text{wall}}} = 0,\ v|_{\Gamma_{\text{in}}} = u_{\text{in}}\} = u^E + \mathcal{V},$$

where u^E is any function that satisfies $u^E|_{\Gamma_{\text{in}}} = u_{\text{in}}$. (The pressure space would have to be modified to $\{q \in L^2(\Omega) \mid \int_\Omega q dx = 0\}$ if we wish to estimate flow in a closed domain without inflow and outflow boundaries, as there is no "absolute" pressure in this case. We will not consider this closed-domain case in this lecture.) We then multiply the momentum equation by $v \in \mathcal{V}$ and the continuity equation by $q \in \mathcal{Q}$, integrate by parts, and incorporate the boundary condition on Γ_{out}. The resulting weak formulation is as follows: for each $t \in (0, T]$, find $(u(t), p(t)) \in \mathcal{V}^E \times \mathcal{Q}$ such that

$$\int_\Omega v \cdot \frac{\partial u}{\partial t}(t) dx + \int_\Omega \nu \nabla v : \nabla u(t) dx + \int_\Omega v \cdot (u(t) \cdot \nabla u(t)) dx - \int_\Omega (\nabla \cdot v) p(t) dx = 0 \quad \forall v \in \mathcal{V},$$
$$- \int_\Omega q (\nabla \cdot u(t)) dx = 0 \quad \forall q \in \mathcal{Q}.$$

We could more compactly express the equations by introducing bilinear forms $m(\cdot, \cdot)$, $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ and a trilinear form $c(\cdot, \cdot, \cdot)$: for each $t \in (0, T]$, find $(u(t), p(t)) \in \mathcal{V}^E \times \mathcal{Q}$ such that

$$m(\frac{\partial u}{\partial t}(t), v) + a(u(t), v) + c(u(t), u(t), v) + b(p(t), v) = 0 \quad \forall v \in \mathcal{V},$$
$$b(q, u(t)) = 0 \quad \forall q \in \mathcal{Q},$$

where

$$
\begin{aligned}
m(w,v) &\equiv \int_\Omega v\cdot w dx = \int_\Omega v_j w_j dx \qquad \forall w,v\in H^1(\Omega)^d,\\
a(w,v) &\equiv \int_\Omega \nu\nabla v : \nabla w dx = \int_\Omega \nu \frac{\partial v_i}{\partial x_j}\frac{\partial w_i}{\partial x_j} dx \qquad \forall w,v\in H^1(\Omega)^d,\\
b(q,v) &\equiv -\int_\Omega (\nabla\cdot v) q dx = -\int_\Omega \frac{\partial v_j}{\partial x_j} q dx \qquad \forall q\in L^2(\Omega),\ \forall v\in H^1(\Omega)^d,\\
c(z,w,v) &\equiv \int_\Omega v\cdot(z\cdot\nabla w) dx = \int_\Omega v_i z_j \frac{\partial w_i}{\partial x_j} dx \qquad \forall z,w,v\in H^1(\Omega)^d.
\end{aligned}
$$

We obtain the *steady* Navier-Stokes equations by setting $\frac{\partial u}{\partial t}=0$: find $(u,p)\in\mathcal{V}^E\times\mathcal{Q}$ such that

$$
\begin{aligned}
a(u,v)+c(u,u,v)+b(p,v) &= 0 \quad \forall v\in\mathcal{V},\\
b(q,u) &= 0 \quad \forall q\in\mathcal{Q}.
\end{aligned}
$$

We will focus on the solution of this steady form of the equations for the rest of the lecture.

13.3 Well-posedness: Stokes problem

The (steady) Navier-Stokes equations, unlike equations that we have studied so far, is neither coercive nor linear. The lack of coercivity and linearity necessitates the use of new technical tools to show well-posedness in the sense that the solution (i) exists, (ii) is unique, and (iii) is stable. In this lecture we introduce tools required to analyze non-coercive problems but not nonlinear problems.

To this end, we introduce the Stokes equations, which models incompressible flows in the limit of vanishing inertia effect relative to the viscous effect (i.e., the Reynolds number approaching 0). The weak form of the Stokes equations is as follows: find $(u,p)\in\mathcal{V}\times\mathcal{Q}$ such that

$$
\begin{aligned}
a(u,v)+b(p,v) &= \ell(v) \quad \forall v\in\mathcal{V},\\
b(q,u) &= 0 \quad \forall q\in\mathcal{Q},
\end{aligned}
\tag{13.1}
$$

where $\ell\in\mathcal{V}'$ is a continuous linear form. Note that nonhomogeneous Dirichlet boundary conditions can be absorbed in the linear form $\ell(\cdot)$, and hence we choose $\mathcal{V}$, and not $\mathcal{V}^E$, as the trial space.

We recall that the Lax-Milgram theorem played a key role in showing the well-posedness of coercive problems. The theorem that plays a similar role for non-coercive problems is the Banach-Nečas-Babuška (BNB) theorem:

Theorem 13.1 (Banach-Nečas-Babuška (BNB) theorem)**.** Let $\mathcal{W}$ be a Banach space, $\mathcal{V}$ be a reflexive Banach space, $A:\mathcal{W}\times\mathcal{V}\to\mathbb{R}$ be a bilinear form, and $\ell\in\mathcal{V}'$ be a linear form. Consider a weak problem: find $u\in\mathcal{W}$ such that $A(u,v)=\ell(v)\ \forall v\in\mathcal{V}$. Then, the weak problem is well-posed if and only if there exists the inf-sup constant

$$
\beta \equiv \inf_{w\in\mathcal{W}}\sup_{v\in\mathcal{V}} \frac{A(w,v)}{\|w\|_{\mathcal{W}}\|v\|_{\mathcal{V}}} > 0 \tag{13.2}
$$

and

$$
\forall v\in\mathcal{V},\quad (\forall w\in\mathcal{W},\ A(w,v)=0)\ \Rightarrow\ (v=0).
$$

In addition, the following stability statement holds:

$$\|u\|_{\mathcal{W}} \leq \frac{1}{\beta}\|\ell\|_{\mathcal{V}'}.$$

Proof. See Ern and Guermond, Theory and Practice of Finite Elements, 2004. □

The BNB theorem provides necessary *and* sufficient condition for a weak formulation to be well-posed. This is unlike the Lax-Milgram theorem, which identifies conditions that are sufficient but not necessary for well-posedness. The BNB theorem is more general than the Lax-Milgram theorem; the latter can be readily derived from the former for coercive problems.

The application of the BNB theorem to the Stokes problem yields the following result:

Proposition 13.2. The Stokes problem (13.1) has a unique solution, and the following stability statement holds:

$$\|u\|_{\mathcal{V}} \leq \frac{1}{\alpha}\|\ell\|_{\mathcal{V}'},$$
$$\|p\|_{\mathcal{Q}} \leq \frac{1}{\beta}\left(1+\frac{\gamma}{\alpha}\right)\|\ell\|_{\mathcal{V}'},$$

where the coercivity and continuity constant for $a(\cdot,\cdot)$ are given by

$$\alpha \equiv \inf_{v\in\mathcal{Z}} \frac{a(v,v)}{\|v\|_{\mathcal{V}}^2} \qquad \text{and} \qquad \gamma \equiv \sup_{w\in\mathcal{V}}\sup_{w\in\mathcal{V}} \frac{|a(w,v)|}{\|w\|_{\mathcal{V}}\|v\|_{\mathcal{V}}}$$

for the space $\mathcal{Z} \equiv \{v \in \mathcal{V} \mid b(q,v) = 0 \ \forall q \in \mathcal{Q}\}$, and the Babuška-Brezzi (BB) inf-sup constant is given by

$$\beta \equiv \inf_{q\in\mathcal{Q}} \sup_{v\in\mathcal{V}} \frac{b(q,v)}{\|q\|_{\mathcal{Q}}\|v\|_{\mathcal{V}}} > 0. \tag{13.3}$$

Sketch of proof. We provide a sketch of a proof. For a complete proof, see, for example, Ern and Guermond, Theory and Practice of Finite Elements, 2004.

We first observe that in the space $\mathcal{Z} \equiv \{v \in \mathcal{V} \mid b(q,v) = 0 \ \forall q \in \mathcal{Q}\}$ the Stokes problem (13.1) simplifies to the following: find $u \in \mathcal{Z}$ such that

$$a(u,v) = \ell(v) \quad \forall v \in \mathcal{Z}. \tag{13.4}$$

We note that the space $\mathcal{Z} \subset \mathcal{V}$ is a Hilbert space. In addition, the bilinear form $a(\cdot,\cdot)$ is coercive in $\mathcal{V}$ because (i) $a(\cdot,\cdot)$ is equivalent to the $H^1(\Omega)$ semi-norm and (ii) we can invoke the Poincaré-Friedrichs inequality. It follows that $a(\cdot,\cdot)$ is also coercive in $\mathcal{Z} \subset \mathcal{V}$. The bilinear form $a(\cdot,\cdot)$ is also continuous in $\mathcal{Z} \subset \mathcal{V}$. Hence, by the Lax-Milgram theorem in $\mathcal{Z}$, the problem (13.4) has a unique solution and satisfies the stability statement $\|u\|_{\mathcal{V}} \leq \frac{1}{\alpha}\|\ell\|_{\mathcal{V}'}$.

We next observe that given velocity $u \in \mathcal{V}$, the pressure satisfies the following statement: find $p \in \mathcal{Q}$ such that

$$b(p,v) = \ell(v) - a(u,v) \quad \forall v \in \mathcal{V}.$$

For $\mathcal{V}$ such that $H_0^1(\Omega)^d \subset \mathcal{V} \subset H^1(\Omega)^d$ and $\mathcal{Q} \equiv L^2(\Omega)$, it can be shown that the BB inf-sup constant β is bounded away from 0. Hence the solution p exists and is unique by the BNB theorem. In addition, by the BNB theorem and continuity of $a(\cdot,\cdot)$ and ℓ,

$$\|p\|_{\mathcal{Q}} \leq \frac{1}{\beta}\|\ell(\cdot) - a(u,\cdot)\|_{\mathcal{V}'} \leq \frac{1}{\beta}(\|\ell\|_{\mathcal{V}'} + \gamma\|u\|_{\mathcal{V}}) \leq \frac{1}{\beta}\left(1+\frac{\gamma}{\alpha}\right)\|\ell\|_{\mathcal{V}'},$$

which is the desired stability result. □

Note that the combination of the coercivity constant α and the inf-sup constant β plays the role of a stability constant in the Stokes problem.

Remark 13.3. The Stokes problem is a prototypical *saddle-point problem*, and admits a saddle-point formulation. To see this, we first introduce a Lagrangian

$$\mathcal{L}(w,q) \equiv \frac{1}{2}a(w,w) + b(q,w) - \ell(w) \quad \forall w \in \mathcal{V},\ \forall q \in \mathcal{Q}.$$

The solution to the Stokes problem (13.1), $(u,p) \in \mathcal{V} \times \mathcal{Q}$, is the saddle point of the Lagrangian in the sense that

$$\mathcal{L}(u,p) = \inf_{w\in\mathcal{V}} \sup_{q\in\mathcal{Q}} \mathcal{L}(w,q).$$

In the saddle-point formulation, the pressure p is the Lagrange multiplier that enforces the divergence-free condition. The well-posed analysis we have considered in this section applies to any saddle-point problem.

13.4 Finite element formulation

We now consider finite element approximation of the Stokes and Navier-Stokes equations. To this end, we first introduce approximation spaces associated with *Taylor-Hood elements*:

$$\mathcal{V}_h = \{v \in \mathcal{V} \mid v|_K \in \mathbb{P}^p(K)^d,\ \forall K \in \mathcal{T}_h\},$$
$$\mathcal{Q}_h = \{q \in \mathcal{Q} \mid q|_K \in \mathbb{P}^{p-1}(K),\ \forall K \in \mathcal{T}_h\}.$$

In words, $\mathcal{V}_h$ is the space of piecewise $\mathbb{P}^p$ polynomials with appropriate Dirichlet boundary conditions, and $\mathcal{Q}_h$ is the space of piecewise $\mathbb{P}^{p-1}$ polynomials. As we will see shortly, as the equations are only inf-sup stable and not coercive, the use of non-equal polynomial degrees for $\mathcal{V}_h$ and $\mathcal{Q}_h$ is necessary to ensure the finite element problem is well posed. For problems with nonhomogeneous Dirichlet boundaries, we also introduce

$$\mathcal{V}_h^E \equiv u_h^E + \mathcal{V}_h,$$

where $u_h^E \in \mathcal{V}_h$ satisfies $u_h^E|_\Gamma = u_h^b$.

The finite element approximation of the Stokes equations is as follows: find $(u_h, p_h) \in \mathcal{V}_h \times \mathcal{Q}_h$ such that

$$a(u_h, v_h) + b(p_h, v_h) = \ell(v_h) \quad \forall v_h \in \mathcal{V}_h,$$
$$b(q_h, u_h) = 0 \quad \forall q_h \in \mathcal{Q}_h.$$

Similarly the finite element approximation of the Navier-Stokes equations is as follows: find $(u_h, p_h) \in \mathcal{V}_h^E \times \mathcal{Q}_h$ such that

$$a(u_h, v_h) + c(u_h, u_h, v_h) + b(p_h, v_h) = 0 \quad \forall v_h \in \mathcal{V}_h,$$
$$b(q_h, u_h) = 0 \quad \forall q_h \in \mathcal{Q}_h.$$

Remark 13.4. Similarly to the advection-diffusion equation considered in Lecture 9, the finite element approximation of the Navier-Stokes equations may exhibit spurious oscillations if the grid Peclet number is large: i.e., $\frac{|u|h_K}{\nu} \gg 1$. The spurious oscillations can be eliminated by incorporating a GLS or SUPG stabilization considered in Lecture 9. In this lecture, we for simplicity assume that h is sufficiently small such that $\frac{|u|h_K}{\nu} < 1$, and we do not consider these stabilization techniques.

13.5 Finite element theory

We now analyze the error in the finite element approximation. We limit our analysis to the Stokes equations. We begin with a generalized Céa's lemma which applies to non-coercive problems that satisfy the BNB conditions:

Lemma 13.5 (Céa's lemma (non-coercive))**.** Consider the setting of the BNB theorem, Theorem 13.1, and the associated finite-dimensional approximation: find $u_h \in \mathcal{W}_h \subset \mathcal{W}$ such that $A(u_h, v_h) = \ell(v_h)\ \forall v_h \in \mathcal{V}_h \subset \mathcal{V}$. Then

$$\|u - u_h\|_{\mathcal{W}} \leq \left(1 + \frac{\gamma}{\beta_h}\right) \inf_{w_h \in \mathcal{W}_h} \|u - w_h\|_{\mathcal{W}},$$

where the (discrete) inf-sup constant β_h and the continuity constant γ are given by

$$\beta_h \equiv \inf_{w_h \in \mathcal{W}_h} \sup_{v_h \in \mathcal{V}_h} \frac{A(w_h, v_h)}{\|w_h\|_{\mathcal{W}} \|v_h\|_{\mathcal{V}}},$$
$$\gamma \equiv \sup_{w \in \mathcal{W}} \sup_{v \in \mathcal{V}} \frac{A(w, v)}{\|w\|_{\mathcal{W}} \|v\|_{\mathcal{V}}}.$$

Proof. By the definition of the discrete inf-sup constant, for an arbitrary $w_h \in \mathcal{W}_h$,

$$\beta_h \|u_h - w_h\|_{\mathcal{W}} \leq \sup_{v_h \in \mathcal{V}_h} \frac{A(u_h - w_h, v_h)}{\|v_h\|_{\mathcal{V}}} = \sup_{v_h \in \mathcal{V}_h} \frac{A(u - w_h, v_h)}{\|v_h\|_{\mathcal{V}}} \leq \gamma \|u - w_h\|_{\mathcal{W}},$$

where the equality follows from the Galerkin orthogonality $A(u - u_h, v_h) = 0\ \forall v_h \in \mathcal{V}_h$, which implies $A(u_h - w_h, v_h) = A(u - w_h, v_h)\ \forall w_h, v_h \in \mathcal{V}_h$. It follows that by the triangle inequality

$$\|u - u_h\|_{\mathcal{W}} \leq \|u - w_h\|_{\mathcal{W}} + \|u_h - w_h\|_{\mathcal{W}} \leq \left(1 + \frac{\gamma}{\beta_h}\right) \|u - w_h\|.$$

We choose w_h to be the infimizer of the norm to conclude the proof. □

We again note that the inf-sup constant β_h plays the role of the "stability constant" that the coercivity constant α plays for coercive problems. However, one key distinction is that that *the discrete inf-sup constant is not bounded from below by the (continuous) inf-sup constant*: i.e.,

$$\beta_h \equiv \inf_{w_h \in \mathcal{W}_h} \sup_{v_h \in \mathcal{V}_h} \frac{A(w_h, v_h)}{\|w_h\|_{\mathcal{W}} \|v_h\|_{\mathcal{V}}} \not\geq \inf_{w \in \mathcal{W}} \sup_{v \in \mathcal{V}} \frac{A(w, v)}{\|w\|_{\mathcal{W}} \|v\|_{\mathcal{V}}} \equiv \beta,$$

because while the infimizer space is smaller the supremizer space is also smaller. This is unlike the case of the coercivity constant for which

$$\alpha_h \equiv \inf_{w_h \in \mathcal{W}_h} \frac{a(w_h, w_h)}{\|w_h\|_{\mathcal{W}}^2} \geq \inf_{w \in \mathcal{W}} \frac{a(w, w)}{\|w\|_{\mathcal{W}}^2} \equiv \alpha.$$

The fact $\beta_h \not\geq \beta$ means that, for non-coercive problems, the approximation of the weak solution in a subspace (as we do in finite element) may not produce a well-posed problem even if the original problem is well-posed. This in turn often implies that we have to either (i) choose the approximation spaces carefully such that the discrete inf-sup condition is satisfied for all h or (ii) add a stabilization term to the bilinear form.

We now apply the generalized Céa's lemma to the saddle-point problem associated with the Stokes equations. To begin, we introduce the discrete Babuška-Brezzi condition associated with Stokes (or more generally saddle-point) problems:

Proposition 13.6 (Babuška-Brezzi condition)**.** A finite-element approximation of the Stokes equation is well-posed if the Babuška-Brezzi condition

$$\beta_h \equiv \inf_{q_h \in \mathcal{Q}_h} \sup_{v_h \in \mathcal{V}_h} \frac{b(q_h, v_h)}{\|q_h\|_{\mathcal{Q}} \|v_h\|_{\mathcal{V}}} > 0$$

is satisfied.

We again observe that the Babuška-Brezzi inf-sup constant β_h is not bounded from below by the continuous counterpart β:

$$\beta_h \equiv \inf_{q_h \in \mathcal{Q}_h} \sup_{v_h \in \mathcal{V}_h} \frac{b(q_h, v_h)}{\|q_h\|_{\mathcal{Q}} \|v_h\|_{\mathcal{V}}} \not\geq \inf_{q \in \mathcal{Q}} \sup_{v \in \mathcal{V}} \frac{b(q, v)}{\|q\|_{\mathcal{Q}} \|v\|_{\mathcal{V}}} \equiv \beta.$$

We cannot choose an infimizer space (i.e., the pressure space $\mathcal{Q}_h$) that is "too large" relative to the supremizer space (i.e., the velocity space $\mathcal{V}_h$) as such a choice could yield $\beta_h = 0$ and an ill-posed finite element problem. In particular, if we choose the equal-degree polynomials for $\mathcal{V}_h$ and $\mathcal{Q}_h$, then the finite element approximation of the Stokes problem is *not* well-posed. The Babuška-Brezzi condition is also referred to as *inf-sup condition*, *Ladyzhenskaya-Babuška-Brezzi (LBB) condition*, or *compatibility condition* in literature.

There are two approaches to ensure that the Babuška-Brezzi condition is satisfied: the first is to choose an appropriate pair of $\mathcal{V}_h$ and $\mathcal{Q}_h$; the second is to add an explicit pressure stabilization. In this lecture we pursue the first approach and use the Taylor-Hood elements, for which the following result holds:

Proposition 13.7 (Taylor-Hood elements)**.** The $\mathbb{P}^p$-$\mathbb{P}^{p-1}$ Taylor-Hood element space, based on piecewise $\mathbb{P}^p$ velocity space and piecewise $\mathbb{P}^{p-1}$ pressure space, satisfies the Babuška-Brezzi condition for any triangulation $\mathcal{T}_h$.

Having ensured the well-posedness of the finite element approximation of the Stokes equation, we next analyze the error in the finite element approximation:

Proposition 13.8. Consider the finite element approximation of the Stokes problem with the velocity approximation space $\mathcal{V}_h \subset \mathcal{V}$ and the pressure approximation space $\mathcal{Q}_h \subset \mathcal{Q}$. Let $\mathcal{Z}_h \subset \mathcal{V}_h$ be the space of discretely divergence-free functions: $\mathcal{Z}_h \equiv \{v_h \in \mathcal{V}_h \mid b(q_h, v_h) = 0, \forall q_h \in \mathcal{Q}_h\}$. We introduce the following constants:

$$\alpha_h \equiv \inf_{v_h \in \mathcal{Z}_h} \frac{a(v_h, v_h)}{\|v_h\|_{\mathcal{V}}^2}, \qquad \gamma \equiv \sup_{w \in \mathcal{V}} \sup_{v \in \mathcal{V}} \frac{a(w, v)}{\|w\|_{\mathcal{V}} \|v\|_{\mathcal{V}}},$$
$$\beta_h \equiv \inf_{q_h \in \mathcal{Q}_h} \sup_{v_h \in \mathcal{V}_h} \frac{b(q_h, v_h)}{\|q_h\|_{\mathcal{Q}} \|v_h\|_{\mathcal{V}}}, \qquad \delta \equiv \sup_{q \in \mathcal{Q}} \sup_{v \in \mathcal{V}} \frac{b(q, v)}{\|q\|_{\mathcal{Q}} \|v\|_{\mathcal{V}}}.$$

The following error estimates hold:

$$\|u - u_h\|_{\mathcal{V}} \leq \left(1 + \frac{\gamma}{\alpha_h}\right) \inf_{w_h \in \mathcal{Z}_h} \|u - w_h\|_{\mathcal{V}} + \frac{\delta}{\alpha_h} \inf_{q_h \in \mathcal{Q}_h} \|p - q_h\|_{\mathcal{Q}},$$
$$\|p - p_h\|_{\mathcal{Q}} \leq \frac{\gamma}{\beta_h} \left(1 + \frac{\gamma}{\alpha_h}\right) \inf_{w_h \in \mathcal{Z}_h} \|u - w_h\|_{\mathcal{V}} + \left(1 + \frac{\delta}{\beta_h} + \frac{\gamma\delta}{\alpha_h \beta_h}\right) \inf_{q_h \in \mathcal{Q}_h} \|p - q_h\|_{\mathcal{Q}}.$$

Moreover, the best-fit approximation error in $\mathcal{Z}_h$ is bounded by

$$\inf_{w_h \in \mathcal{Z}_h} \|u - w_h\|_{\mathcal{V}} \leq \left(1 + \frac{\delta}{\beta_h}\right) \inf_{v_h \in \mathcal{V}_h} \|u - v_h\|_{\mathcal{V}}.$$

Proof. See Ern and Guermond, Theory and Practice of Finite Elements, 2010. □

In words, the Taylor-Hood approximation of the Stokes problem is quasi-optimal in the sense that the finite-element approximation is only some constant away from the best-fit approximation for a given space $\mathcal{V}_h \times \mathcal{Q}_h$ The combination of the proposition and the polynomial approximation theory yields the following result:

Proposition 13.9. If $u \in H^1(\Omega) \cap H^{s+1}(\mathcal{T}_h)$ and $p \in L^2(\Omega) \cap H^{s'}(\mathcal{T}_h)$, the error in the finite element approximation based on the $\mathbb{P}^p$-$\mathbb{P}^{p-1}$ Taylor-Hood elements is bounded by

$$\|u - u_h\|_{H^1(\Omega)} \leq C_1 h^r (|u|_{H^{r+1}(\mathcal{T}_h)} + |p|_{H^r(\mathcal{T}_h)})$$
$$\|p - p_h\|_{L^2(\Omega)} \leq C_2 h^r (|u|_{H^{r+1}(\mathcal{T}_h)} + |p|_{H^r(\mathcal{T}_h)})$$

where $r \equiv \min\{s, s', p\}$.

The proposition shows that if the solution is smooth then the finite element approximation converges as $\|u - u_h\|_{H^1(\Omega)} \leq Ch^p$ and $\|p - p_h\|_{L^2(\Omega)} \leq C'h^p$. In addition, if the problem is sufficiently regular such that the elliptic regularity estimate hold, then $\|u - u_h\|_{L^2(\Omega)} \leq C''h^{p+1}$.

13.6 Finite element implementation

We now consider the implementation of the finite element method. To this end, we first introduce polynomial spaces

$$H^1_{h,p}(\Omega) \equiv \{v \in C^1(\overline{\Omega}) \mid v|_K \in \mathbb{P}^p(K),\ \forall K \in \mathcal{T}_h\},$$
$$H^1_{h,p-1}(\Omega) \equiv \{v \in C^1(\overline{\Omega}) \mid v|_K \in \mathbb{P}^{p-1}(K),\ \forall K \in \mathcal{T}_h\},$$

which do not incorporate any essential boundary conditions. We then introduce bases $\{\phi_k\}_{k=1}^m$ and $\{\chi_{k'}\}_{k'=1}^{m'}$ so that

$$H^1_{h,p}(\Omega) = \text{span}\{\phi_k\}_{k=1}^m \quad \text{and} \quad H^1_{h,p-1}(\Omega) = \text{span}\{\chi_{k'}\}_{k=1}^{m'}.$$

We can then uniquely express any vector-valued velocity $u_h \in H^1_{h,p}(\Omega)^d$ and any scalar-valued pressure $p_h \in H^1_{h,p-1}(\Omega)$ as

$$u_{h,j}(x) = \sum_{k=1}^m \hat{u}_{h,jk}\phi_k(x) \quad \text{and} \quad p_h(x) = \sum_{k'=1}^{m'} \hat{p}_{h,k'}\chi_{k'}(x)$$

for some coefficients $\hat{u}_h \in \mathbb{R}^{dm}$ and $\hat{p}_h \in \mathbb{R}^{m'}$. For notational convenience, we also introduce the "full" coefficient vector $\overline{U} \in \mathbb{R}^{dm+m'}$ such that

$$\overline{U} = \begin{pmatrix} \hat{u}_{h,1} \\ \vdots \\ \hat{u}_{h,d} \\ \hat{p}_h \end{pmatrix}.$$

Our goal is to find the vector $\overline{U} \in \mathbb{R}^{dm+m'}$ associated with the finite element approximation.

To define the solution, we first define the (discrete) residual operator associated with our finite element approximation that does not (yet) incorporate the Dirichlet boundary conditions: $\overline{R} : \mathbb{R}^{dm+m'} \to \mathbb{R}^{dm+m'}$ given by

$$\overline{R}(\overline{U}) = \begin{pmatrix} \overline{R}_1(\overline{U}) \\ \vdots \\ \overline{R}_d(\overline{U}) \\ \overline{R}_{d+1}(\overline{U}) \end{pmatrix}$$

where

$$\begin{aligned} \overline{R}_i(\overline{U})_a &= \int_\Omega \left(\nu \frac{\partial \phi_a}{\partial x_k} \frac{\partial u_{h,i}}{\partial x_k} + \phi_a u_{h,k} \frac{\partial u_{h,i}}{\partial x_k} - \frac{\partial \phi_a}{\partial x_i} p_h \right) dx, \quad i = 1, \ldots, d, \\ \overline{R}_{d+1}(\overline{U})_{a'} &= -\int_\Omega \chi_{a'} \left(\frac{\partial u_{h,j}}{\partial x_j} \right) dx, \end{aligned}$$

for $a = 1, \ldots, m$ and $a' = 1, \ldots, m'$. The first d blocks are associated with the momentum equation, and the last block is associated with the continuity equation. For instance, for a two-dimensional problems the residual is given by

$$\begin{aligned} \overline{R}_1(\overline{U})_a &= \int_\Omega \left(\nu \frac{\partial \phi_a}{\partial x_1} \frac{\partial u_{h,1}}{\partial x_1} + \nu \frac{\partial \phi_a}{\partial x_2} \frac{\partial u_{h,1}}{\partial x_2} + \phi_a u_{h,1} \frac{\partial u_{h,1}}{\partial x_1} + \phi_a u_{h,2} \frac{\partial u_{h,1}}{\partial x_2} - \frac{\partial \phi_a}{\partial x_1} p_h \right) dx \\ \overline{R}_2(\overline{U})_a &= \int_\Omega \left(\nu \frac{\partial \phi_a}{\partial x_1} \frac{\partial u_{h,2}}{\partial x_1} + \nu \frac{\partial \phi_a}{\partial x_2} \frac{\partial u_{h,2}}{\partial x_2} + \phi_a u_{h,1} \frac{\partial u_{h,2}}{\partial x_1} + \phi_a u_{h,2} \frac{\partial u_{h,2}}{\partial x_2} - \frac{\partial \phi_a}{\partial x_2} p_h \right) dx \\ \overline{R}_3(\overline{U})_{a'} &= -\int_\Omega \chi_{a'} \left(\frac{\partial u_{h,1}}{\partial x_1} + \frac{\partial u_{h,2}}{\partial x_2} \right) dx, \end{aligned}$$

for $a = 1, \ldots, m$ and $a' = 1, \ldots, m'$.

We now consider two distinct approaches to impose Dirichlet boundary condition that yield exactly the same solution. The first approach is consistent with our function space interpretation $\mathcal{V}_h \subset H^1_{h,p}(\Omega)^d$; however, this approach is admittedly more complicated to implement than the second approach and hence is *not* recommended. We first decompose the coefficient vector $\overline{U}$ as

$$\overline{U} = U^e + \tilde{U},$$

where $U^e \in \mathbb{R}^{dm+m'}$ is chosen such that the associated finite element solution (u^e_h, p^e_h) satisfies — or more precisely approximates — the Dirichlet boundary conditions (i.e., the values of U^e on a Dirichlet boundary node is given by the associated boundary value), and $\tilde{U} \in \mathbb{R}^{dm+m'}$ is a vector that is zero on Dirichlet boundary nodes. We next introduce a vector $\hat{R} \in \mathbb{R}^{dn+n'}$ which we obtain by removing the degrees of freedom associated with the Dirichlet boundary nodes from $\overline{R} \in \mathbb{R}^{dm+m'}$. The solution to our finite element problem is given by the following equations for the coefficients: find $\overline{U} = U^e + \tilde{U} \in \mathbb{R}^{dm+m'}$ such that

$$\hat{R}(\overline{U}) = 0 \quad \text{in } \mathbb{R}^{dn+n'}.$$

This approach is consistent with our construction $\mathcal{V}^E = u^E + \mathcal{V}_h$, but is somewhat cumbersome to implement.

Alternative — implementationally simpler and recommended — approach to impose the Dirichlet boundary condition is to replace the residual associated with Dirichlet nodes with a "penalty equation" associated with the boundary condition as follows:

$$\overline{R}_i^E(\overline{U})_a = \begin{cases} \overline{R}_i(\overline{U})_a & \text{if component } i \text{ of node } a \text{ is not Dirichlet BC} \\ \hat{u}_{h,ia} - u_i^b(x_a) & \text{if component } i \text{ of node } a \text{ is Dirichlet BC} \end{cases},$$

for $i = 1, \ldots, d$. This approach is simpler to implement as (i) the size of the residual vector is unchanged and (ii) the Dirichlet boundary condition is imposed as part of the residual.

The solution to our finite element problem is given by the following equations for the coefficients: find $\bar{U} \in \mathbb{R}^{dm+m'}$ such that

$$\bar{R}^E(\bar{U}) = 0 \quad \text{in } \mathbb{R}^{dm+m'}.$$

This equation is nonlinear in $\bar{U}$; we will find the solution using Newton's method.

13.7 Solution of nonlinear problems by Newton's method

We now solve the nonlinear equation for the coefficients $\overline{U}$ using Newton's method. To this end, we first identify the Jacobian matrix $\overline{J}(\overline{U}) \in \mathbb{R}^{(dm+m')\times(dm+m')}$ given by

$$\overline{J}(\overline{U}) = \frac{\partial \overline{R}}{\partial \overline{U}} = \begin{pmatrix} \overline{J}_{1,1}(\overline{U}) & \cdots & \overline{J}_{1,d}(\overline{U}) & \overline{J}_{1,d+1}(\overline{U}) \\ \vdots & \ddots & \vdots & \vdots \\ \overline{J}_{d,1}(\overline{U}) & \cdots & \overline{J}_{d,d}(\overline{U}) & \overline{J}_{d,d+1}(\overline{U}) \\ \overline{J}_{d+1,1}(\overline{U}) & \cdots & \overline{J}_{d+1,d}(\overline{U}) & 0 \end{pmatrix},$$

where

$$\begin{aligned}
\overline{J}_{i,j}(\overline{U})_{ab} &= \left[\frac{\partial \overline{R}_i}{\partial \overline{U}_j}\right]_{a,b} = \int_\Omega \left(\nu \frac{\partial \phi_a}{\partial x_k}\frac{\partial \phi_b}{\partial x_k}\delta_{ij} + \phi_a u_{h,k}\frac{\partial \phi_b}{\partial x_k}\delta_{ij} + \phi_a \frac{\partial u_{h,i}}{\partial x_j}\phi_b\right) dx, \quad i,j = 1,\ldots,d, \\
\overline{J}_{i,d+1}(\overline{U})_{ab'} &= \left[\frac{\partial \overline{R}_i}{\partial \overline{U}_{d+1}}\right]_{a,b'} = -\int_\Omega \frac{\partial \phi_a}{\partial x_i}\chi_{b'} dx, \quad i = 1,\ldots,d, \\
\overline{J}_{d+1,j}(\overline{U})_{a'b} &= \left[\frac{\partial \overline{R}_{d+1}}{\partial \overline{U}_j}\right]_{a',b} = -\int_\Omega \chi_{a'}\frac{\partial \phi_a}{\partial x_j} dx, \quad j = 1,\ldots,d,
\end{aligned}$$

for $a, b = 1, \ldots, m$ and $a', b' = 1, \ldots, m'$. For instance, for a two-dimensional problems the Jacobian is given by

$$\overline{J}(\overline{U}) = \frac{\partial \overline{R}}{\partial \overline{U}} = \begin{pmatrix} \overline{J}_{11}(\overline{U}) & \overline{J}_{12}(\overline{U}) & \overline{J}_{13}(\overline{U}) \\ \overline{J}_{21}(\overline{U}) & \overline{J}_{22}(\overline{U}) & \overline{J}_{23}(\overline{U}) \\ \overline{J}_{31}(\overline{U}) & \overline{J}_{32}(\overline{U}) & 0 \end{pmatrix},$$

where

$$\overline{J}_{11}(\overline{U})_{ab} = \left[\frac{\partial \overline{R}_1}{\partial \overline{U}_1}\right]_{a,b} = \int_\Omega \left(\nu \frac{\partial \phi_a}{\partial x_1}\frac{\partial \phi_b}{\partial x_1} + \nu \frac{\partial \phi_a}{\partial x_2}\frac{\partial \phi_b}{\partial x_2} + \phi_a u_{h,1}\frac{\partial \phi_b}{\partial x_1} + \phi_a u_{h,2}\frac{\partial \phi_b}{\partial x_2} + \phi_a \phi_b \frac{\partial u_{h,1}}{\partial x_1}\right) dx$$

$$\overline{J}_{12}(\overline{U})_{ab} = \left[\frac{\partial \overline{R}_1}{\partial \overline{U}_2}\right]_{a,b} = \int_\Omega \phi_a \frac{\partial u_{h,1}}{\partial x_2}\phi_b dx,$$

$$\overline{J}_{13}(\overline{U})_{ab'} = \left[\frac{\partial \overline{R}_1}{\partial \overline{U}_3}\right]_{a,b'} = -\int_\Omega \frac{\partial \phi_a}{\partial x_1}\chi_{b'} dx,$$

$$\overline{J}_{21}(\overline{U})_{ab} = \left[\frac{\partial \overline{R}_2}{\partial \overline{U}_1}\right]_{a,b} = \int_\Omega \phi_a \frac{\partial u_{h,2}}{\partial x_1}\phi_b dx,$$

$$\overline{J}_{22}(\overline{U})_{ab} = \left[\frac{\partial \overline{R}_2}{\partial \overline{U}_2}\right]_{a,b} = \int_\Omega \left(\nu \frac{\partial \phi_a}{\partial x_1}\frac{\partial \phi_b}{\partial x_1} + \nu \frac{\partial \phi_a}{\partial x_2}\frac{\partial \phi_b}{\partial x_2} + \phi_a u_{h,1}\frac{\partial \phi_b}{\partial x_1} + \phi_a u_{h,2}\frac{\partial \phi_b}{\partial x_2} + \phi_a \phi_b \frac{\partial u_{h,2}}{\partial x_2}\right) dx,$$

$$\overline{J}_{23}(\overline{U})_{ab'} = \left[\frac{\partial \overline{R}_2}{\partial \overline{U}_3}\right]_{a,b'} = -\int_\Omega \frac{\partial \phi_a}{\partial x_2}\chi_{b'} dx,$$

$$\overline{J}_{31}(\overline{U})_{a'b} = \left[\frac{\partial \overline{R}_3}{\partial \overline{U}_1}\right]_{a,b'} = -\int_\Omega \chi_{a'}\frac{\partial \phi_b}{\partial x_1} dx,$$

$$\overline{J}_{32}(\overline{U})_{a'b} = \left[\frac{\partial \overline{R}_3}{\partial \overline{U}_2}\right]_{a,b'} = -\int_\Omega \chi_{a'}\frac{\partial \phi_b}{\partial x_2} dx,$$

for $a, b = 1, \ldots, m$ and $a', b' = 1, \ldots, m'$.

We now impose Dirichlet boundary conditions. We recall that we have considered two distinct approaches to implement Dirichlet boundary conditions. For the first approach, we introduce $\hat{J} \in \mathbb{R}^{(dn+n')\times(dn+n')}$ which we obtain by removing the degrees of freedom associated with the Dirichlet boundary nodes from $\overline{J} \in \mathbb{R}^{(dm+m')\times(dm+m')}$. Note that this operation involves the elimination of both the rows and columns of the Jacobian matrix associated with Dirichlet nodes.

For the second (and again simpler and recommended) approach to impose Dirichlet boundary conditions, we consider the Jacobian $\overline{J}^E(\overline{U})$ associated with $\overline{R}^E(\cdot)$ whose entries are given by

$$\overline{J}^E_{ij}(\overline{U})_{a,b} = \begin{cases} \overline{J}_{ij}(\overline{U})_{a,b} & \text{if } a \text{ is not a Dirichlet node,} \\ 1 & \text{if component } i \text{ of node } a \text{ is Dirichlet BC and } a = b \text{ and } i = j, \\ 0 & \text{if component } i \text{ of node } a \text{ is Dirichlet BC and } a \neq b \text{ or } i \neq j, \end{cases}$$

for $i, j = 1, \ldots, d$. This operation corresponds to setting all entries of the rows associated with Dirichlet nodes equal to 0 except for the diagonal entry (i.e., $a = b$ and $i = j$) which is set to 1. Note that we do not modify the columns associated with the Dirichlet node.

Given the expression for the residual $\overline{R}^E(\overline{U})$ and Jacobian $\overline{J}^E(\overline{U})$, we solve for the root of $\overline{R}^E(\cdot)$ using Newton's method. Here, for simplicity, we consider the second (and recommend) approach to imposing Dirichlet boundary conditions. Newton's method proceeds as follows:

0. Initialize state coefficient vector $\overline{U}^{k=0} \in \mathbb{R}^{dm+m'}$. Set $k = 0$.

1. Evaluate the residual and Jacobian

$$\overline{R}^E(\overline{U}^k) \in \mathbb{R}^{dm+m'} \quad \text{and} \quad \overline{J}^E(\overline{U}^k) \in \mathbb{R}^{(dm+m')\times(dm+m')},$$

where the rows associated with Dirichlet nodes are modified to incorporate the boundary condition.

2. If $\|\overline{R}^E(\overline{U}^k)\|_\infty \leq \epsilon_{\text{tol}}$, then terminate.

3. Compute the update $\delta\overline{U}^k \in \mathbb{R}^{dm+m'}$ by solving

$$[\overline{J}^E(\overline{U}^k)]\delta\overline{U}^k = -\overline{R}^E(\overline{U}^k) \quad \text{in } \mathbb{R}^{dm+m'}.$$

4. Update the state as $\overline{U}^{k+1} = \overline{U}^k + \delta\overline{U}^k$.

5. Set $k \leftarrow k+1$, and return to 1.

We obtain a solution $\overline{U}$ which satisfies the residual as well as Dirichlet boundary conditions at convergence.

Remark 13.10. If the problem is strongly nonlinear and the initial state is far from the solution, then (pure) Newton's method as described here might not converge. In these cases, Newton's method must be used in conjunction with a *homotopy* (or *continuation*) strategy. The idea is to solve a sequence of increasingly nonlinear problems, the last of which is the problem of interest. In the case of Navier-Stokes equations, we can first solve the problem for a large kinematic viscosity ν, which yields a nearly linear problem with a Stokes-like solution, and then successively decrease ν (i.e., increase the Reynolds number) using the solution for the previous (higher) ν case as the initial state for the new (lower) ν case.

13.8 Variational Newton's method

We can also describe the Newton's method in any (infinite-dimensional) Hilbert space. To this end, we first introduce the residual form

$$r((u,p),(v,q)) \equiv a(u,v) + c(u,u,v) + b(p,v) + b(q,u) \quad \forall u,v \in \mathcal{V},\ \forall p,q \in \mathcal{Q}.$$

(The spaces $\mathcal{V}$ and $\mathcal{Q}$ can be replaced the finite-dimensional counterpart $\mathcal{V}_h \subset \mathcal{V}$ and $\mathcal{Q}_h \subset \mathcal{Q}$.) The associated Fréchet derivative is

$$\begin{aligned} r'((u,p);(w,z),(v,q)) &\equiv \lim_{\epsilon\to 0} \frac{1}{\epsilon}\left[r((u+\epsilon w, p+\epsilon z),(v,q)) + r((u,p),(v,q))\right] \\ &= a(w,v) + c(z,u,w) + c(u,w,v) + b(z,v) + b(q,w). \end{aligned}$$

The variational Newton's method proceeds as follows:

0. Initialize state $(u^k,p^k) \in \mathcal{V}\times\mathcal{Q}$. Set $k=0$.

1. Terminate if

$$\|r((u^k,p^k),(\cdot,\cdot))\|_{(\mathcal{V}\times\mathcal{Q})'} = \sup_{(v,q)\in\mathcal{V}\times\mathcal{Q}} |r((u^k,p^k),(v,q))| \leq \epsilon_{\text{tol}}.$$

2. Evaluate the update: find $(\delta u^k, \delta p^k) \in \mathcal{V}\times\mathcal{Q}$ such that

$$r'((u^k,p^k);(\delta u^k,\delta p^k),(v,q)) = -r((u^k,p^k),(v,q)) \quad \forall v\in\mathcal{V},\ \forall q\in\mathcal{Q}.$$

3. Update the state $u^{k+1} = u^k + \delta u^k$ and $p^{k+1} = p^k + \delta p^k$.

4. Set $k \leftarrow k + 1$, and return to 1.

The procedure at termination yields a solution $(u, p) \in \mathcal{V} \times \mathcal{Q}$ that satisfies the Navier-Stokes equations.

13.9 Summary

We summarize key points of this lecture:

1. Incompressible flows are modeled by the Navier-Stokes equations, which is a system of nonlinear PDEs with vector-valued velocity and scalar-valued pressure.

2. Incompressible flows in the limit of vanishing inertia effect relative to viscosity is modeled by the Stokes equations, which is a linear saddle-point problem.

3. The weak formulation of the Stokes and Navier-Stokes equations are defined for a velocity space $\mathcal{V} \subset H^1(\Omega)^d$ and a pressure space $\mathcal{Q} \subset L^2(\Omega)$. The inertia term of the Navier-Stokes equations involve quadratic nonlinearity, which can be concisely expressed as a trilinear form.

4. The Stokes problem is well-posed: it has a unique solution and is stable.

5. The finite element approximation is well-posed if the velocity and pressure spaces are chosen to satisfy the Babuška-Brezzi condition (or ins-sup condition). The $\mathbb{P}^p$-$\mathbb{P}^{p-1}$ Taylor-Hoods elements are an example of finite elements that satisfy the Babuška-Brezzi condition uniformly in h.

6. The $\mathbb{P}^p$-$\mathbb{P}^{p-1}$ Taylor-Hoods elements provide a quasi-optimal approximation for both the velocity and pressure.

7. The solution to the nonlinear algebraic equation associated with the Navier-Stokes equations is obtained using Newton's method. The method requires the evaluation of the residual and the associated Jacobian.

8. Newton's method can also be described in any (infinite-dimensional) Hilbert space.