Finite Element Analysis

Adrian J. Lew
Gustavo C. Buscaglia

June 6, 2023

Contents

Chapter 1

Finite Element Methods for Elliptic Problems in 1D

1.1 Second-Order Problems

1.1.1 The Differential Equation

Linear, second-order elliptic differential equations in one dimension have the general form

$$-\big(k(x)\,u'(x)\big)' + b(x)\,u'(x) + c(x)\,u(x) = f(x)\,, \tag{1.1}$$

where x is the **independent variable**, typically a space variable, k, b, c and f are the **coefficients** of the equation, which may depend on x, and u is the relevant function being studied.

A differential equation is a condition for the function u that must be fulfilled at every point of the domain. At each $x \in \Omega$, the function u must **(a)** be smooth enough for all terms in the equation to be computable and **(b)** satisfy the algebraic equation (1.1). In this case, we say that the function u is a **solution** of the differential equation 1.1.

A multitude of physical problems can be modeled with Equation (1.1). We see some of them next, together with examples of solutions u to 1.1 in particular cases.

Examples:

1.1 Consider a vertical cylindrical column of uniform cross sectional area and height H, Young modulus $E(x)$ and density $\rho(x)$, x being the vertical coordinate. Then the vertical displacement $u(x)$ of the cross section at height x must satisfy the equilibrium equation

$$-\big(E(x)\,u'(x)\big)' = -\rho(x)\,g\,, \tag{1.2}$$

where g is the magnitude of the acceleration of gravity. Clearly, it reduces to (1.1) by taking $k = E$, $b = 0$, $c = 0$ and $f = -\rho\,g$. The vertical

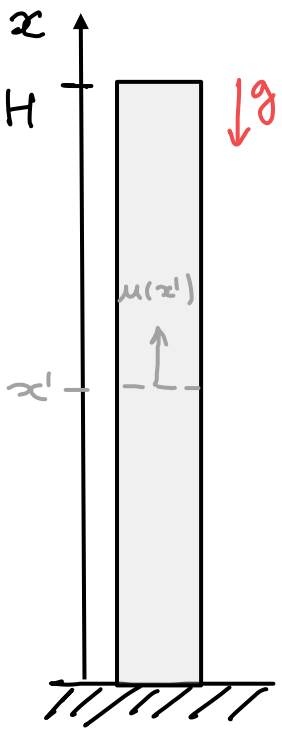


Figure 1.1 Sketch for Example 1.1.

stress (force per unit area) on the cross-section is given by

$$\sigma = E\,u' \,. \tag{1.3}$$

1.2 When heat is flowing through a wall, x being the through-the-wall coordinate, the temperature u obeys the **diffusion equation**

$$-\big(k(x)\,u'(x)\big)' = f(x)\,, \tag{1.4}$$

where $k(x)$ is the **thermal conductivity** of the wall material at point x and f is a volumetric heat source (due for example to γ-radiation).

In the case of a homogeneous wall without volumetric sources, the temperature is affine in x. The temperature gradient, and thus also the **heat flux** $-ku'$, are constant.

1.3 **Steady-state convection-diffusion-reaction equation:** Let $u(x)$ denote $C(x) - C_{\text{eq}}$, where $C(x)$ the concentration of a species in a mobile one-dimensional reactive medium that moves with **velocity** $b(x)$ and C_{eq} is the equilibrium value. Then $u(x)$ satisfies (1.1), with k being the **diffusion coefficient**, c the **reaction coefficient** and f a possible volumetric source of the species.

In this case the first term is called **diffusion term** and consists of the spatial derivative of the **diffusive flux**

$$J_{\text{diff}} = -k\,u' \,.$$

It expresses the differential mass balance due to molecular diffusion. The second term is called **advection term** (sometimes **convection term**). It models the mass balance due to the transport of the species by the movement of the ambient medium. In fact, it is the *material time derivative* of the concentration in steady-state conditions (when the concentration does not depend on time), namely

$$\frac{Du}{Dt} = bu'.$$

In fact, the **advective flux** is

$$J_{\text{adv}} = b\,u \,.$$

The term $c(x)\,u(x)$ models the reaction of the species with the medium towards equilibrium. The coefficient c is proportional to the kinetic constant of the reaction. As c grows the local concentration is increasingly driven towards equilibrium and thus u gets closer to zero.

1.4 If we consider now a horizontal membrane under tension which is subject to vertical loads f, the vertical displacement u satisfies

$$-T\,u''(x) = f(x)\,, \tag{1.5}$$

where T is the **membrane tension**. This equation models the equilibrium of vertical forces on each arbitrary part of the membrane, which in this one-dimensional setting should better be visualized as an elastic string.

We hope the reader is by now convinced of the usefulness of models governed by (1.1). Of course the actual field $u(x)$ that arises for given coefficients (k, b, c and f) depends on the **boundary conditions**. In one dimension, the domain of analysis Ω is usually an interval, which we take as $0 < x < L$, i.e.,

$$\Omega = (0, L) .$$

In the physical systems considered in the previous examples, as in many other models of mathematical physics, the relevant field u arises as the solution of a **boundary value problem**. This means that *one* additional condition is imposed at $x = 0$ and *a second one* at $x = L$, and that these two conditions are necessary and sufficient to fully determine u. This is an intrinsic property of **second-order elliptic problems**: To fully define the solution u of the differential equation, it is necessary and sufficient to specify one condition at all points of the boundary. If there is some part of the boundary where no condition is specified, then there exist infinitely many solutions. In the one-dimensional setting considered here the boundary (denoted in general by $\partial\Omega$) consists of just the extreme points of the interval, i.e.

$$\partial\Omega = \{0, L\} .$$

We will also refer to the **closure** of Ω, defined for this domain as

$$\overline{\Omega} = \Omega \cup \partial\Omega = [0, L].$$

The most popular boundary conditions are

- the **Dirichlet condition**, which imposes the value of u (for example, $u(0) = g_0$ or $u(L) = g_L$),
- and the **Neumann condition**, which imposes the value of u' (for example $u'(0) = d_0$ or $u'(L) = d_L$).

The problem that we introduce now has a Dirichlet condition at $x = 0$ and a Neumann condition at $x = L$. Other possibilities of boundary conditions will be discussed later.

Problem 1.1. *Given the coefficients k, b, c and f (as functions of x), together with the boundary constants g_0 and d_L, find a continuous function $u : \Omega \to \mathbb{R}$ satisfying*

$$-\big(k(x)u'(x)\big)' + b(x)u'(x) + c(x)u(x) = f(x) \qquad \forall\, x \in \Omega \tag{1.6a}$$

$$u(0) = g_0 \tag{1.6b}$$

$$u'(L) = d_L \tag{1.6c}$$

Examples:

1.5 Consider a purely diffusive ($b = c = 0$), homogeneous ($k(x) = k_0,\ \forall\, x$) case without source ($f = 0$). The constants g_0 and d_L remain arbitrary. Then the solution of the problem in strong form must be continuous and satisfy

$$-u''(x) = 0, \quad \forall x \in (0, L), \qquad u(0) = g_0, \qquad \text{and } \ u'(L) = d_L$$

Polynomials of degree ≤ 1, i.e., of the form $c_1 + c_2 x$, have vanishing second derivative in $(0, L)$, and they are continuous functions of x. They are therefore solutions of the differential equation. Further, choosing $c_1 = g_0$ and $c_2 = d_L$ we identify the only polynomial solution to Problem 1.1, namely

$$u(x) = g_0 + d_L\, x\,.$$

Further, it is known that the *only* functions that have zero second derivative in an interval are polynomials of degree 1. Thus the function $u(x)$ above is the *unique* solution to Problem 1.1.

1.6 **Problems without solution.** Problem 1.1 does not always have a solution. Consider a case with constant diffusivity ($k = 1$), with no convection or source ($b = f = 0$) and with $c(x) = 1/x^2$. So, the equation reads

$$-u''(x) + \frac{u(x)}{x^2} = 0, \qquad \forall x \in (0, L)\,. \tag{1.7}$$

It can be checked by substitution that any function of the form

$$u(x) = c_1\, x^{(1+\sqrt{5})/2} + c_2\, x^{(1-\sqrt{5})/2}$$

satisfies (1.7). In fact, it is known that these functions are the *only* solutions of (1.7). However, notice that the exponent of x in the second term is *negative*, so that if $c_2 \neq 0$ the solution is $\pm\infty$ at $x = 0$ and thus different from g_0. So, for the problem to have a solution, $c_2 = 0$. But now, since the exponent of x in the first term is *positive*, the value of u at $x = 0$ is zero for any value of c_1! Unless the given value of g_0 is zero, there is no solution. Just for completeness, in the particular case $g_0 = 0$ there is indeed a unique solution, and the constant c_1 can be computed so that $u'(L) = d_L$.

1.7 **Deformed Column.** Consider again Example 1.1 of a vertical cylindrical column of uniform cross sectional area, height H, Young modulus $E(x)$ and density $\rho(x)$, with x being the vertical coordinate, see Fig. 1.1. Assume that the column is unloaded on its top face, and supported on a rigid foundation at its base. These define the boundary conditions of the problem by

$$u(0) = 0, \qquad \sigma(H) = E(H)\,u'(H) = 0,$$

which are of the Dirichlet and Neumann type, respectively. These equations, together with (1.2), define the strong form of this problem.
To solve it, we integrate (1.3) over a slice of the column, from $x = h_1$ to $x = h_2$ to get

$$\sigma(h_1) - \sigma(h_2) = -g \int_{h_1}^{h_2} \rho(x)\, dx$$

so that the difference in σ between two positions equals the weight of the slice (per unit area). Since the column is unloaded on its top face, $\sigma(H) = 0$, and hence

$$\sigma(x) = -g \int_{x}^{H} \rho(\xi)\, d\xi.$$

Therefore, from (1.3), the displacement of the column follows as

$$u(x) - u(0) = \int_0^x \frac{\sigma(\xi)}{E(\xi)}\, d\xi = -\int_0^x \frac{g}{E(\xi)} \int_{\xi}^{H} \rho(y)\, dy. \tag{1.8}$$

Since the foundation is rigid, $u(0) = 0$. We can verify next that function $u(x)$ defined in (1.8) is a solution of the differential equation (1.2). First, we can compute $(E(x)u'(x))'$ for any point x by using the fundamental lemma of calculus, so it is smooth enough for all terms in the equation to be computable (condition (**a**) above), and these terms satisfy the algebraic equation (1.2) at any point x.
In the particular case in which $E(x) = E$ and $\rho(x) = \rho$, both constants through the length of the column, the solution u is

$$u(x) = -x(2H - x)\frac{g\rho}{2E}.$$

What about the solution of Problem 1.1? Does it exist at all? Is it unique? The answer to this question is derived from the general theory of linear ordinary differential equations, and the answer is **yes**, but of course conditional to some hypotheses. The hypotheses we consider here correspond to *elliptic* problems, and as we have seen allow us to model a wide variety of physical problems.

In general, sound physical models lead to well-posed mathematical problems, that is, problems for which a unique solution exists, and the solution changes smoothly when the coefficients of the equation or the boundary conditions do. However, theorems are helpful references to go to in case of doubt. We state here an existence and uniqueness theorem that covers most applications.

Theorem 1.1 (Existence and Uniqueness of Solutions)**.** *Assume that $k(x)$, $b(x)$, $c(x)$ and $f(x)$ are smooth and bounded, and also that $k(x) \geq k_0 > 0$. Further, let $c_0 = \min_x c(x)$ and assume that $c_0 \geq 0$. Then Problem 1.1 has a unique solution.*

This is more than what we need to know at this point about the strong form of the elliptic second-order boundary-value problem that we are set to analyze.

1.1.2 Variational Equations

The finite element method is based on the observation that the solution of Problem 1.1 satisfies a **variational equation**. A variational equation is not only a staple of finite element methods, but it is always a puzzling and welcome surprise to those who are introduced to it for the first time. Let's see what a variational equation for Problem 1.1 looks like.

Let u be the solution of Problem 1.1. Then, u satisfies that

$$\int_\Omega \left[k(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)\right]\,dx \\ - k(L)d_L v(L) = \int_\Omega f(x)v(x)\,dx. \quad (1.9a)$$

for any $v \in \mathcal{V}$, where

$$\mathcal{V} = \{w\colon [0,L] \to \mathbb{R} \text{ smooth } \mid w(0) = 0\}. \quad (1.9b)$$

The *variational equation* is (1.9a), while (1.9b) defines the **test space** $\mathcal{V}$. A functions $v \in \mathcal{V}$ is called a **test function**. The name variational equation originates in the appearance of equations of this form in the Calculus of Variations, or the general theory of extreme values (see, e.g., [10]).

For now, we will understand a function to be **smooth** in the definition of $\mathcal{V}$ as a function in which all derivatives exists and are continuous in $[0,L]$. Later on, we will expand this definition to include more functions, so that some commonly used functions used in the finite element method are also included in $\mathcal{V}$.

Equation (1.9a) is called a variational equation because it has the form

$$F(u,v) = 0 \qquad \forall v \in \mathcal{V}, \quad (1.10)$$

for a scalar-valued function F and a test space $\mathcal{V}$. In our example,

$$F(u,v) = \int_\Omega \left[k(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)\right]\,dx \\ - k(L)d_L v(L) - \int_\Omega f(x)v(x)\,dx.$$

We will have the opportunity to more precisely define what a variational equation is later.

Notice that a variational equation is a statement of a potentially infinite set of equations, since each function in $\mathcal{V}$ defines a condition that u needs to satisfy. If there are enough functions in $\mathcal{V}$, or enough conditions, a variational equation may be able to define u as the unique function that can satisfy it.

Example 1.8 Let's illustrate that variational equation (1.9a) is satisfied for a solution of Problem (1.1) and some choice of test function.

To this end, let $\Omega = (0,1)$, $k(x) = b(x) = c(x) = 1$ and $f(x) = -5\exp(-2x)$ for all $x \in \Omega$, $g_0 = 1$, and $d_L = -2\exp(-2)$. Equations (1.6) from the strong form of Problem 1.1 become

$$-u''(x) + u'(x) + u(x) = -5\exp(-2x), \qquad \forall\, x \in \Omega, \tag{1.11a}$$

$$u(0) = 1, \tag{1.11b}$$

$$u'(1) = -2\exp(-2). \tag{1.11c}$$

The exact solution of this problem is $u(x) = \exp(-2x)$.

As a function in the test space, consider the function $v(x) = \sin(x)$, which satisfies that $v(0) = 0$, so $v \in \mathcal{V}$. The left hand side of variational equation (1.9a) then reads

$$\int_0^1 (\exp(-2x))'\sin'(x) + (\exp(-2x))'\sin(x) + \exp(-2x)\sin(x)\, dx + 2\exp(-2)\sin(1) =$$

$$\int_0^1 -2\exp(-2x)\cos(x) - 2\exp(-2x)\sin(x) + \exp(-2x)\sin(x)\, dx + 2\exp(-2)\sin(1) =$$

$$\cos(1)\exp(-2) - 1 + 2\exp(-2)\sin(1).$$

The right hand side of the same equation is

$$-\int_0^1 5\exp(-2x)\sin(x)\ \ dx \ = \ -1 + \cos(x)\exp(-2) + 2\sin(1)\exp(-2).$$

Since both sides have the same value, (1.9a) is satisfied. A similar result would follow for any $v \in \mathcal{V}$.

Instead, if $v \notin V$, (1.9a) may not be satisfied. For example, if $v(x) = \cos(x)$, then the left hand side equals $\left[1 + \exp(-2)\,(9\cos(1) - 2\sin(1))\right]/5$, and the right hand side differs from it with the value $-1 + \exp(-2)\,(\cos(1) + 2\sin(1))$.

1.1.2.1 Derivation of a Variational Equation

How do we obtain a variational equation that the solution of Problem 1.1 satisfies? To illustrate it, we begin with the simplest possible case, and set $k(x) = 1$, $b(x) = c(x) = 0$ for all $x \in \Omega = (0, L)$, so that

$$-u''(x) = f(x) \qquad x \in \Omega, \tag{1.12a}$$

$$u(0) = g_0, \tag{1.12b}$$

$$u'(L) = d_L. \tag{1.12c}$$

The variational equation is obtained from Problem 1.1 by following three steps:

1. Multiply the partial differential equation (1.6a) by an arbitrary smooth function $v\colon \Omega \to \mathbb{R}$ and integrate over the interval $[0, L]$ to obtain

$$0 = u''(x)\,v(x) + f(x)\,v(x) \Rightarrow 0 = \int_0^L u''(x)\,v(x) + f(x)\,v(x)\,dx \tag{1.13}$$

2. Integrate by parts[1] the second derivative of $u''(x)$, to pass the derivative to v, to get:

$$0 = u'(L)\,v(L) - u'(0)\,v(0) - \int_0^L u'(x)\,v'(x)\,dx + \int_0^L f(x)\,v(x)\,dx. \tag{1.14}$$

3. Use boundary condition (1.12c) to replace the value of $u'(L) = d_L$ in (1.14), and since we know nothing about the value of $u'(0)$, we will require $v(0) = 0$. Thus,

$$0 = \underbrace{u'(L)}_{=d_L,\ \text{due to (1.12}b\text{)}}\ v(L) - u'(0)\ \underbrace{v(0)}_{=0,\ \text{require it from } v\in\mathcal{V}} - \int_0^L u'(x)\,v'(x)\,dx + \int_0^L f(x)\,v(x)\,dx, \tag{1.15}$$

   from where it follows that

$$\int_0^L u'(x)\,v'(x)\,dx - d_L\,v(L) = \int_0^L f(x)\,v(x)\,dx. \tag{1.16}$$

   for any $v \in \mathcal{V}$, where

$$\mathcal{V} = \{w\colon [0, L] \to \mathbb{R} \text{ smooth } | \, v(0) = 0\}.$$

Equation (1.16) is a variational equation that the solution u of Problem 1.1 satisfies. It is the particularization of (1.9a) for the values of k, b, c and Ω in this example.

Thus, in obtaining variational equation (1.16) we also introduced conditions that test functions $v \in \mathcal{V}$ should satisfy. What if we did not require $v(0) = 0$ for $v \in \mathcal{V}$? In this case, we would need to retain the second term on the left hand side in step 3 above, (1.15), and would have obtained a different variational equation that u satisfies, namely,

$$\int_0^L u'(x)\,v'(x)\,dx + u'(0)\,v(0) - d_L\,v(L) = \int_0^L f(x)\,v(x)\,dx. \tag{1.17}$$

[1] For two smooth functions w and v, $(wv)' = w'v + wv'$ by the product formula, and integrating on both sides we get the integration by parts formula:

$$\lim_{x\to b^-} w(x)\,v(x) - \lim_{x\to a^+} w(x)\,v(x) = \int_a^b w'(x)\,v(x)\,dx + \int_a^b w(x)\,v'(x)\,dx.$$

We use it by setting $w = u'$.

for any $v \in \mathcal{V}$, where

$$\mathcal{V} = \{w\colon [0, L] \to \mathbb{R} \text{ smooth }\}.$$

We may go even further and question what if we did not integrate by parts in step 2 above. In that case, we obtain yet another variational equation that u satisfies (from (1.13)):

$$0 = \int_0^L u''(x)\, v(x) + f(x)\, v(x)\, dx \tag{1.18}$$

for any $v \in \mathcal{V}$, where

$$\mathcal{V} = \{w\colon [0, L] \to \mathbb{R}\}, \tag{1.19}$$

where we do not require smoothness for functions in $\mathcal{V}$ since we are not integrating by parts and computing derivatives of test functions; we just need the integrals we are performing to be computable.

This discussion shows that u satisfies different variational equations, each one for a potentially different set of test functions. In fact, there are an infinite number of variational equations that u satisfies. The most remarkable difference between the variational equations we listed is the way that u, v and their derivatives are involved, and the boundary conditions required from functions in the test space. In (1.16), only the first derivative of u and v in Ω is needed; in (1.14) we additionally need the value of the first derivative of u on part of the boundary of Ω, but no boundary conditions on test functions; and in (1.18) two derivatives of u are required, while no derivative of test functions are needed, nor boundary conditions on them.

Each variational equation could serve as the basis for the formulation of a finite element method. For reasons that we will have the chance to discuss later, the most commonly adopted variational equation is (1.16), because it has the same number of derivatives required from both u and v, and no evaluation of derivatives of u on the boundary of the domain.

1.1.2.2 Essential and Natural Boundary Conditions

In obtaining variational equation (1.16) we incorporated boundary condition (1.12c) into it when we replaced $u'(L)$ by d_L. In contrast, boundary condition $u(0) = g_0$, (1.12b), was not used at all. Because of this, *any* function that satisfies partial differential equation (1.12a) and boundary condition (1.12c) also satisfies variational equation (1.16), even if they do not satisfy boundary condition (1.12b).

Examples:

1.9 In (1.12), let $L = \pi/2$, $f(x) = \cos(x)$, $g_0 = 1$ and $d_L = -1$. Then, the function $u(x) = \cos(x)$ is the solution of (1.12). As such, it satisfies variational equation (1.16) for any smooth v such that $v(0) = 0$. For

example, for $v(x) = \sin(x)$,

$$\underbrace{\int_0^{\pi/2} -\sin(x)\cos(x)\,dx}_{=-1/2} + 1\sin(\pi/2) - \underbrace{\int_0^{\pi/2} \cos(x)\sin(x)\,dx}_{=-1/2} = 0.$$

Consider then $u_1(x) = u(x) + 1 = \cos(x) + 1$, which satisfies (1.12a) and (1.12c), but does not satisfy (1.12b), since $u_1(0) = 2 \neq g_0$. In spite of this, function u_1 also satisfies variational equation (1.16), as it can be easily inferred from the fact that $u' = u_1'$.

In contrast, consider $\hat{u}_2(x) = u(x) + x = \cos(x) + x$, which satisfies partial differential equation (1.12a) but *not* boundary condition (1.12c), since $u_2'(\pi/2) = 0 \neq -1$. This function does not satisfy variational equation (1.16): for $v(x) = \sin(x)$,

$$\underbrace{\int_0^{\pi/2} (1-\sin(x))\cos(x)\,dx}_{=1/2} + 1\sin(\pi/2) - \underbrace{\int_0^{\pi/2} \cos(x)\sin(x)\,dx}_{=-1/2} = 1 \neq 0.$$

1.10 For $f(x) = \cos(x)$ and $L = \pi/2$, the general solution of the differential equation (1.12a) is $u(x) = c_1 + c_2 x + \cos(x)$, for any $c_1, c_2 \in \mathbb{R}$. The values of the two constants c_1, c_2 can be determined so as to satisfy the boundary conditions (1.12b) and (1.12c), respectively. In fact,

$$u(0) = c_1 + 1, \qquad \text{and} \qquad u'(\pi/2) = c_2 - 1.$$

Because we are discussing about variational equations, we will see now that we can determine the value of c_2 from variational equation (1.16) by selecting a suitable test function v. In other words, we will see that the variational equation defines the value of the Neumann boundary condition.

As a first case, let's choose $v \in \mathcal{V}$ such that $v(L) = v(\pi/2) = 0$. For example, we can choose $v(x) = x(x - \pi/2)$, a quadratic polynomial that has zeros at $x = 0$ and $x = \pi/2$. Replacing in the left hand side of the variational equation, we obtain

$$\int_0^{\pi/2} (c_2 - \sin(x))(2x - \pi/2)\,dx = (\pi - 4)/2.$$

Computing the right hand side we obtain the same value, namely,

$$\int_0^{\pi/2} \cos(x)x(x - \pi/2)\,dx = (\pi - 4)/2,$$

so the variational equation is satisfied for this choice of $v \in \mathcal{V}$, *regardless* of the values of c_1 and c_2. This choice of test function does not add any condition to the two free constants.

Instead, let's choose $v(x) = x$ as a test function, which is different than zero at $x = \pi/2$; $v(\pi/2) = \pi/2$. In this case, the left hand side evaluates to

$$\int_0^{\pi/2} (c_2 - \sin(x))\, dx - d_L \pi/2 = \frac{\pi}{2}(c_2 - d_L) - 1.$$

The right hand side takes the value

$$\int_0^{\pi/2} \cos(x) x\, dx = \pi/2 - 1.$$

Equating both sides, we obtain an equation that involves c_2, namely,

$$\frac{\pi}{2}(c_2 - d_L) - 1 = \pi/2 - 1,$$

and solving it for c_2 we conclude that $c_2 = 1 + d_L$.

In summary, by testing with $v(x) = x(x - L)$ the variational equation did not impose any condition on the values of c_1 and c_2. However, by testing with another element of $\mathcal{V}$, $v(x) = x$, we obtained a necessary condition for the variational equation to be satisfied: $c_2 = 1 + d_L$.

This means that not all solutions of (1.12a) satisfy variational equation (1.16). A necessary condition is that c_2 acquire a specific value. Therefore, the general solution of (1.12a) that could satisfy the variational equation is

$$u(x) = c_1 + (1 + d_L)x + \cos(x).$$

Since $u'(\pi/2) = 1 + d_L - \sin(\pi/2) = d_L$, u satisfies the Neumann boundary condition (1.12c) for any value of c_1.

These examples illustrate the more general fact that for a function to satisfy a variational equation, it may need to satisfy some boundary conditions as well. It also illustrates that not all boundary conditions in a problem may need to be satisfied for the solution to satisfy a variational equation.

Therefore, given a variational equation that the solution of a problem such as Problem 1.1 needs to satisfy, we can classify the boundary conditions of the problem into two types:

Natural Boundary Conditions (NBC): Boundary conditions that any function that satisfies the variational equation needs to satisfy.

Essential Boundary Conditions (EBC): Any boundary condition of the problem that is not a natural boundary condition.

If we derive the variational equation, as we did earlier, then all boundary conditions that we incorporate into it during the derivation will be natural boundary conditions. All other boundary conditions will be essential. Instead, if we are given a variational equation, we will learn later how to infer what natural boundary conditions it requires.

Examples:

1.11 For variational equation (1.16), $u(0) = g_0$ is an essential boundary condition, and $u'(L) = d_L$ is a natural boundary condition. This is also true for variational equation (1.17).

1.12 For variational equation (1.18), both $u(0) = g_0$ and $u'(L) = d_L$ are essential boundary condition, since none of them was incorporated in the variational equation during its derivation.

1.1.2.3 A Recipe to Obtain Variational Equations

In the following we describe a recipe to obtain variational equations from a partial differential equation and boundary conditions. We illustrate each step of the recipe with Problem 1.1 and derive (1.93), but the recipe works for a large class of problems. The steps of the recipe are:

1. **Form the residual:** Begin by forming the residual of (1.6a): subtract the right hand side from the left hand side of the equality (or vice versa). That is, for a function u we define a function $r\colon [0,L] \to \mathbb{R}$ as

$$r = -(k(x)u'(x))' + b(x)u'(x) + c(x)u(x) - f(x). \tag{1.20}$$

   Then, according to (1.6a), for the solution u of the strong form we should have

$$r(x) = 0 \qquad x \in (0,L). \tag{1.21}$$

2. **Multiply by a test function and integrate.** We then proceed and multiply this equation by a function $v \in \mathcal{V}$ and integrate over $(0,L)$, where $\mathcal{V}$ is some set of smooth functions over $(0,L)$ that we shall specify later. As aforementioned, functions $v \in \mathcal{V}$ are called *test* functions, but are also labeled **weight** functions. For any such $v \in \mathcal{V}$, we have

$$\int_0^L r(x)\,v(x)\,dx = 0. \tag{1.22}$$

   Again, we are replacing the requirement $r(x) = 0$ for all $x \in (0,L)$, for (1.22) to be satisfied for all functions in $\mathcal{V}$. Because the residual functions are multiplied by the weight functions, this form of formulating the problem is also called the *Method of Weighted Residuals* (MWR)[2].

   In our example, this means:

$$\int_0^L \big(-(k(x)u'(x))' + b(x)u'(x) + c(x)u(x) - f(x)\big)\,v\,dx = 0 \tag{1.23}$$

   for all $v \in \mathcal{V}$.

3. **Integrate the residual by parts.** Assuming that both u and v are smooth enough for the integration by parts formula to hold, integrate by parts terms in the residual $m \geq 0$ times, with m less or equal than the maximum number of derivatives of u in the residual. For each value of m a *different* variational equation and hence weak form is obtained. Typically, we integrate by parts until the number of derivatives of u is equal or only one higher than the number of derivatives of v in each term. In other words, use integration by parts to "transfer" derivatives from u to v, until u has either an equal number of or one more derivative than v in any term in the resulting expression. In our example, $m = 1$, and this leads to

$$\int_0^L k(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x) - f(x)v(x)\,dx \\ - k(L)u'(L)v(L) + k(0)u'(0)v(0) = 0 \quad (1.24)$$

   In this example, we only integrated by parts the first term of the residual, since the second term already has the desired difference in the order of the derivatives between u and v.

4. **Use boundary conditions and identify conditions for $\mathcal{V}$.** Out of the boundary terms that appear from integrating by parts, identify those for which the value has been provided or can be solved for from the boundary conditions, and replace them in the boundary terms. As with the integration by parts, there could be some ambiguity here, leading to different weak formulations. In this case, (1.6c) gives the value of $u'(L)$. Replacing in our example,

$$\int_0^L k(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x) - f(x)v(x)\,dx \\ - k(L)d_L v(L) + k(0)u'(0)v(0) = 0 \quad (1.25)$$

   However, we do not know anything about the value of $u'(0)$; we only know about $u(0)$. In this case, we request the value of the accompanying test function to be zero through the definition of $\mathcal{V}$. In a general case, we proceed similarly with any other boundary term for which we do not have any boundary condition, and require the value of the accompanying (derivative of the) test function to be zero in the definition of $\mathcal{V}$. This is the process to identify conditions that we need to impose for functions in $\mathcal{V}$.

   In our example, we are going to request that if $v \in \mathcal{V}$, then $v(0) = 0$. For any such v,

$$\int_0^L k(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x) - f(x)v(x)\,dx \\ - k(L)d_L v(L) = 0 \quad (1.26)$$

   Those boundary conditions that we are able to incorporate into the variational equation by replacing some of the boundary terms are the natural

boundary conditions for the problem. The remaining boundary conditions need to be explicitly requested for u to satisfy; they are the essential boundary conditions. Hence, for our example, boundary condition (1.12b) is an essential boundary condition.

If we do not request $v(0) = 0$ in the definition of $\mathcal{V}$, we would arrive to a *different* variational equation, still satisfied by the exact solution, with a different set of essential and natural boundary conditions. Similarly, we would obtain an alternative variational equation for each number of derivatives we decide to transfer from u to v.

5. **State the variational equation.** For our example, this leads to (1.93), namely, u satisfies that

$$\int_\Omega \left[k(x)u'(x)v'(x) + b(x)u(x)'v(x) + c(x)u(x)v(x) \right] dx - k(L)d_L v(L) = \int_\Omega f(x)v(x)\, dx. \quad (1.27a)$$

for any $v \in \mathcal{V}$, where

$$\mathcal{V} = \{w : \Omega \to \mathbb{R} \text{ smooth } \mid w(0) = 0\}. \quad (1.27b)$$

Also, list the essential boundary conditions of the problem, since these will be used later on when we formulate a finite element method.

As remarked in step 3, it would also be possible to choose a different distribution of derivatives between u and v than the guideline provided above, leading to a different variational equation. For example, we could transfer all derivatives to v, or leave all derivatives in u. In finite element analysis, we are generally interested in variational equations with minimal smoothness requirements for u and v, since it is simpler to build spaces to approximate solutions in this case.

Example 1.13 A variational equation for a third-order problem. Given $f : [a, b] \to \mathbb{R}$, find $u : [a, b] \to \mathbb{R}$ such that

$$u_{,xxx} = f \qquad x \in (a, b) \quad (1.28a)$$
$$u(a) = 1 \quad (1.28b)$$
$$u_{,x}(b) = 2 \quad (1.28c)$$
$$u_{,xx}(a) = 3. \quad (1.28d)$$

Here the notation $u_{,x}$ denotes the derivative, $u_{,xx}$ the second derivative, etc. The exact solution of this problem is obtained by repeated integration of

(1.28a):

$$\begin{aligned}
\int_a^x f(y)\,dy &= \int_a^x u_{,yyy}(y)\,dy \\
&= u_{,xx}(x) - u_{,xx}(a) = u_{,xx}(x) - 3 \\
\int_b^x \left[\int_a^z f(y)\,dy\right] dz &= \int_b^x u_{,zz}(z) - 3\,dz \\
&= u_{,x}(x) - u_{,x}(b) - 3(x-b) \\
&= u_{,x}(x) - 2 - 3(x-b) \\
\int_a^x \left[\int_b^w \left[\int_a^z f(y)\,dy\right] dz\right] dw &= \int_a^x u_{,w}(w) - 2 - 3(w-b)\,dw \\
&= u(x) - \underbrace{u(a)}_{=1} - 2(x-a) - \frac{3}{2}(x^2-a^2) \\
&\quad + 3b(x-a).
\end{aligned}$$

Therefore,

$$u(x) = 1 + (2-3b)(x-a) + \frac{3}{2}(x^2-a^2) + \int_a^x \left[\int_b^w \left[\int_a^z f(y)\,dy\right] dz\right] dw. \tag{1.29}$$

To identify the variational equation, we proceed as above:

(a) *Form the residual*:

$$r = u_{,xxx} - f.$$

(b) *Multiply by a test function and integrate*:

$$\int_a^b (u_{,xxx} - f)\,v\,dx = 0$$

for all v that is smooth enough.

(c) *Integrate the residual by parts*: In this case, we will integrate by parts only once,

$$u_{,xx}(b)\,v(b) - u_{,xx}(a)\,v(a) - \int_a^b u_{,xx}\,v_{,x} + f\,v\,dx = 0$$

for all v that is smooth enough.

(d) *Use boundary conditions and identify conditions for $\mathcal{V}$*: Here we know the value of $u_{,xx}(a)$, so we need to request $v(b) = 0$. For such v we have

$$-3v(a) - \int_a^b u_{,xx}\,v_{,x} + f\,v\,dx = 0$$

Then, $u_{,xx}(a) = 3$ is a natural boundary condition.

(e) *State the variational equation*: The solution u of (1.28) satisfies that

$$-\int_a^b u_{,xx}\, v_{,x}\, dx = \int_a^b f v\, dx - 3v(a).$$

for all $v \in \mathcal{V}$, where

$$\mathcal{V} = \{v\colon [a,b] \to \mathbb{R} \text{ smooth } \mid v(b) = 0\}.$$

The essential boundary conditions are then $u(a) = 1$ and $u_{,x}(b) = 2$.

1.1.2.4 Other Variational Equations

Not all variational equations follow from the recipe in §1.1.2.3. For example, if a function u satisfies variational equations

$$F(u,v) = 0, \qquad G(u,v) = 0$$

for all $v \in \mathcal{V}$ for scalar-valued functions F and G, then u satisfies the variational equation

$$\alpha F(u,v) + \beta G(u,v) = 0$$

for all $v \in \mathcal{V}$, for any $\alpha, \beta \in \mathbb{R}$.

This enables us to construct a variety of variational equations. Each variational equation could give rise to a different finite element method, as we will have the opportunity to see later. With this in mind, the following are examples of variational equations that give rise to popular finite element methods. We label each example of a variational equation with the method it finds use for.

Example 1.14 Nitsche's Method. A solution u of Problem 1.1 satisfies the following variational equations that can be obtained after multiplying the Dirichlet boundary condition (1.12b) by the value of a test function or its derivative on the boundary

$$(g_0 - u(0))\, v'(0) = 0 \tag{1.30a}$$

$$\mu(u(0) - g_0)\, v(0) = 0 \tag{1.30b}$$

for all $v \in \mathcal{V} = \{v\colon \Omega \to \mathbb{R} \text{ smooth }\}$, where $\mu > 0$ is a positive real number.

If we add the two equations in (1.30) to variational equation (1.17), which has the same test space $\mathcal{V}$, we obtain the following variational equation that u also satisfies:

$$\int_0^L u'(x)v'(x)\, dx + u'(0)v(0) - u(0)v'(0) + \mu u(0)v(0) = \int_0^L f(x)v(x)\, dx + d_L v(L) - g_0 v'(0) + \mu g_0 v(0) \tag{1.31}$$

for all $v \in \mathcal{V} = \{v\colon [0,L] \to \mathbb{R} \text{ smooth}\}$. This variational equation is used to formulate the so-called Nitsche's finite element method for Problem 1.1, in which the Dirichlet boundary condition (1.12b) for u is *also a natural boundary condition* of the problem. There are no essential boundary conditions. We will have the opportunity to discuss this last part in detail in §XXX.

This is a general technique that can be used in a variety of problems. When it is used, we say that we use Nitsche's method for such problem.

Example 1.15 Residual-Stabilized Method. The solution u of Problem 1.1 satisfies the following variational equation

$$0 = \int_{\Omega_E} (u''(x) - f(x))\, v(x)\, dx \tag{1.32}$$

for all $v \in \mathcal{V}_E = \{v\colon \Omega \to \mathbb{R}\}$, where $\Omega_E \subset \Omega$ is a subset of Ω. For example, $\Omega_E = (L/4, L/2)$. When $\Omega_E = \Omega$, we recover (1.18).

We can combine this variational equation with (1.16), namely,

$$\int_0^L u'(x)\, v'(x)\, dx - d_L\, v(L) = \int_0^L f(x)\, v(x)\, dx \tag{1.33}$$

for all $v \in \mathcal{V} = \{v\colon \Omega \to \mathbb{R} \text{ smooth} \mid v(0) = 0\}$, to get that u should satisfy the variational equation

$$\int_0^L u'(x)\, v'(x)\, dx + \mu \int_{\Omega_E} (u''(x) - f(x))\, v(x)\, dx - d_L\, v(L) = \int_0^L f(x)\, v(x)\, dx \tag{1.34}$$

for any $v \in \mathcal{V}$ and some $\mu > 0$. Notice that the test spaces for variational equations (1.32) and (1.33) are different. However, $\mathcal{V} \subset \mathcal{V}_E$, so (1.32) is in particular valid for all $v \in \mathcal{V}$. This is why we can combine them to form (1.34). This variational equation can be used to formulate some of the so-called residual-stabilized finite element methods.

In this case, the Dirichlet boundary condition (1.12b) is an essential boundary condition, and the Neumann boundary condition (1.12c) is a natural one.

Example 1.16 Interior Penalty Methods. The following variational equation is stated on a domain that is split into two or more parts. For simplicity, let's split $\Omega = \Omega_1 \cup \Omega_2$, where for this example, $\Omega_1 = (0, L/2)$ and $\Omega_2 = (L/2, L)$.

The solution u of Problem 1.1 is continuous across the boundary of the two domains, in this case at $x = L/2$, and hence it satisfies that

$$0 = [\![u]\!]|_{x=L/2} = \lim_{x\to L/2^-} u(x) - \lim_{x\to L/2^+} u(x) = u^-(L/2) - u^+(L/2), \qquad (1.35)$$

which just states that the "jump" $[\![u]\!]|_{x=L/2}$ of u at $x = L/2$ should be equal to zero, since the function is continuous there. For convenience, we also introduced the notation $u^+(x) = \lim_{y\to x^+} u(y)$ and $u^-(x) = \lim_{y\to x^-} u(y)$.

For this variational equation, we introduce the test space

$$\mathcal{V} = \{v\colon \Omega \to \mathbb{R} \mid v(0) = 0,\ v \text{ is smooth in } \Omega_1 \text{ and } v \text{ is smooth in } \Omega_2\}.$$

Functions in $\mathcal{V}$ can be discontinuous at $x = L/2$, the common boundary between Ω_1 and Ω_2. Then, the solution u of Problem (1.1) satisfies the following variational equations

$$-(u'v)^-(L/2) + \int_0^{L/2} u'(x)v'(x) = \int_0^{L/2} f(x)v(x)\,dx \qquad (1.36a)$$

$$-d_L v(L) + (u'v)^+(L/2) + \int_{L/2}^{L} u'(x)v'(x) = \int_{L/2}^{L} f(x)v(x)\,dx \qquad (1.36b)$$

$$[\![u]\!]|_{x=L/2}[\![v]\!]|_{x=L/2} = 0 \qquad (1.36c)$$

for all $v \in \mathcal{V}$. The first and second equations, (1.36a) and (1.36b), are obtained by multiplying by v, integrating by parts in Ω_1 and Ω_2, respectively, and using that $v(0) = 0$ and the Neumann boundary condition 1.12c to simplify two of the boundary terms that appear. The third equation, (1.36c), is obtained by multiplying (1.35) by the jump of v at $x = L/2$.

We will combine the three variational equations in (1.38) to form the one we are interested in. But before doing that, it is convenient to state a useful identity. To this end, we define the "average" of a function u at a point x_0 as

$$\{u\}|_{x=x_0} = \frac{u^+(x_0) + u^-(x_0)}{2}.$$

With it, a simple algebraic manipulation of the right hand side of the next identity shows that

$$(u'v)^-(L/2) - (u'v)^+(L/2) = \left([\![u']\!]\{v\} + \{u'\}[\![v]\!]\right)_{x=L/2}. \qquad (1.37)$$

Finally, we obtain a variational equation of the type used in Interior Penalty finite element methods by adding (1.36a), (1.36b) and $\mu > 0$ times (1.36c), and replacing the two boundary terms at $x = L/2$ with (1.37), to get

$$\sum_{i=1}^{2}\int_{\Omega_i} u'(x)v'(x)\,dx - d_L v(L) - ([\![u]\!]\{v'\} + \{u'\}[\![v]\!] - \mu[\![u]\!][\![v]\!])|_{x=L/2} = \sum_{i=1}^{2}\int_{\Omega_i} f(x)v(x)\,dx \qquad (1.38)$$

for all $v \in \mathcal{V}$.

For this variational equation, the Dirichlet boundary condition (1.12b) is an essential boundary condition, while the Neumann boundary condition (1.12c) is a natural boundary condition. As will have the opportunity to discuss in §XXX, this variational equation imposes the condition $[\![u]\!]|_{x=L/2}$, so it is possible to think about the continuity of u as a natural boundary condition for this problem, and not require u to be continuous a priori. Interior Penalty finite element methods are an example of the so-called Discontinuous Galerkin methods, characterized by imposing the continuity of the solution as a natural boundary condition, as in this example.

1.1.3 Sets of Functions*

Part of starting a variational equation is defining a set of test functions for which we require an equation such as (1.9a) to hold (the statement "for all functions v"). Let's describe some common sets of functions and the notation we use to specify them. We proceed by examining some examples.

Examples:

1.17 The set $C^0(I)$ is the set of continuous scalar(real)-valued functions over the interval $I \subset \mathbb{R}$. For example:

i. Let $f(x) = \sin x$. Then, if $I = [0,1]$ we have that $f \in C^0([0,1])$, since $\sin x$ assigns a real value to each point in the interval $[0,1]$, and f is continuous over that interval. Moreover, we have that $f \in C^0(\mathbb{R})$, in which we set $I = \mathbb{R}$, since $\sin x$ is continuous over the entire real line.

ii. The *Heaviside step function* is defined as

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

for $x \in \mathbb{R}$. Then, $H \in C^0((0,1))$ because it is continuous in the open interval $(0,1)$, but $H \notin C^0([-1,1])$, because H is discontinuous at $x = 0$.

1.18 The set $C^k(I)$, for $k \in \mathbb{N}$, is the set of continuous functions with k continuous derivatives over the interval $I \subset \mathbb{R}$.
The set $C^\infty(I)$ is the set of functions in which *all* derivatives are continuous. For example:

i. Let $f(x) = \cos x$. Then, $f \in C^2(\mathbb{R})$ since it has two continuous derivatives anywhere in the real line. Moreover, $f \in C^\infty(\mathbb{R})$, since all derivatives of f are continuous.

ii. Let $g(x) = |x|$, the absolute value function, whose derivative exists anywhere but at $x = 0$. Then, $g \in C^2((0,1])$, $g \in C^0([-1,1])$, but $g \notin C^1([-1,1])$ since g' is discontinuous at $x = 0$, i.e., $\lim_{x\to 0^-} g'(x) \neq \lim_{x\to 0^+} g'(x)$.

1.19 The set $\mathbb{P}_k(I)$, for $k \in \mathbb{N} \cup \{0\}$, is the set of all polynomials of degree less or equal than k over the interval $I \subset \mathbb{R}$. For example:

i. Let $f(x) = x^3 + 1$, then $f \in \mathbb{P}_k(\mathbb{R})$ for any $k \geq 3$.

ii. Let $f(x) = (x-2)^{10}$, then $f \in \mathbb{P}_{10}([0,1])$ for any $k \geq 10$.

A set of functions often contains an infinite number of functions, and it is impossible to enumerate all members of the set. Nevertheless, it is possible to test whether a given function belongs to the set, as we did in the above examples. Additionally, sets of functions are often defined by imposing additional conditions for a function to belong to a set. For example, we could define a set by writing

$$V_1 = \left\{ f \in C^0([0,1] \mid f(0) = 2 \right\},$$

which indicates the set of all continuous functions over the interval $[0,1]$ whose value at $x = 0$ is 2.

We introduced new notation here, which we proceed to explain: The curly brackets $\{\cdot\}$ indicate that what is inside describes the members of the set, and the separator "|" should be read as "*such that.*" So, if we write $V = \{f \in C^0([0,1])\}$ we are saying that the set contains all functions f that are in $C^0([0,1])$; f stands for a generic member of the set. It is equivalent to writing $V = C^0([0,1])$. The expression that defines V_1 above should be read as "*all functions f in $C^0([0,1])$ such that $f(0) = 2$.*" The "|" serves the function of allowing us to add conditions for a function to belong to a set, and we do so by indicating the conditions on the generic member of the set f.

Examples:

1.20 Let $f(x) = x^2$ and $g(x) = x^2 + 2$. Then $f \notin V_1$ and $g \in V_1$.

1.21 Let $V_2 = \left\{ g \in C^2([-1,1]) \mid g(-1) = 1, g'(1) = 2 \right\}$. Then, $x^2 \in V_2$ but $h(x) = x^4 \notin V_2$, since $h'(1) = 4 \neq 2$.

1.22 Let $V_3 = \left\{ h \in C^2([0,L]) \mid h(0) = 0, h(L) = 0 \right\}$. Then $V_3 \subset C^2([0,L])$, that is, the set $C^2([0,L])$ contains all functions in V_3. This is a trivial statement, since in the definition of V_3 we are requesting functions to be in $C^2([0,L])$ as one of the conditions they should satisfy to belong to V_3. However, in the next section we will use the idea that the set $C^2([0,L])$ contains all functions in V_3, so it is a good idea to become familiar with this now.

1.23 Let $V_4 = \{h\colon [a,b] \to \mathbb{R} \text{ smooth} \mid h(a) = 0\}$. In this example, functions in V_4 take a real value for each point in the interval $[a,b]$, and the word

smooth indicates that they should have as many continuous derivatives as required by the problem or the manipulation of expressions we perform; we will talk about what this precisely means later. For example, $\sin(x-a) \in V_4$, since all derivatives exists and are continuous, but the membership of $|x-a|$ will depend on the specifics of the problem.

1.1.4 Integration by Parts of Piecewise Smooth Functions*

Up to now we have been looking at examples in which all functions and a their derivatives are continuous, and we used the integration-by-parts formula on these functions to obtain variational equations. There are incentives, however, to expand the class of functions we consider so as to include functions in which either the function or some of its relevant derivatives are discontinuous. In particular, the finite element method provides a way to construct sets of functions, and the less continuity requirements functions in a set have, the easier it is to construct the set of finite element functions. This is particularly true in two and three spatial dimensions, and over domains that cannot smoothly be mapped to a cube (curved domains, domains with holes, etc.). The integration by parts formula needs to be modified for functions with discontinuities, and this is the focus of the forthcoming discussion.

The type of functions we want to consider are illustrated by the following two (see Fig. 1.2):

- The hat function $N\colon \mathbb{R} \to \mathbb{R}$,

$$N(x) = \begin{cases} 1-|x| & |x| < 1 \\ 0 & |x| \geq 1. \end{cases} \tag{1.39}$$

- The function $M\colon \mathbb{R} \to \mathbb{R}$

$$M(x) = N(x) + \begin{cases} 1 & x \geq 0 \\ 0 & x < 0. \end{cases} \tag{1.40}$$

The hat function N is continuous, while M is not continuous at $x = 0$. They have the same derivative[2], it is

$$N'(x) = M'(x) = \begin{cases} 0 & x < -1 \\ 1 & -1 < x < 0 \\ -1 & 0 < x < 1 \\ 0 & 1 < x \end{cases}$$

[2]Precisely, they have the same **classical derivative**. For a function $N\colon \mathbb{R} \to \mathbb{R}$, the classical derivative at a point x is computed as

$$N'(x) = \lim_{h \to 0} \frac{N(x+h) - N(x)}{h}. \tag{1.41}$$

The classical derivative is defined wherever this limit is.

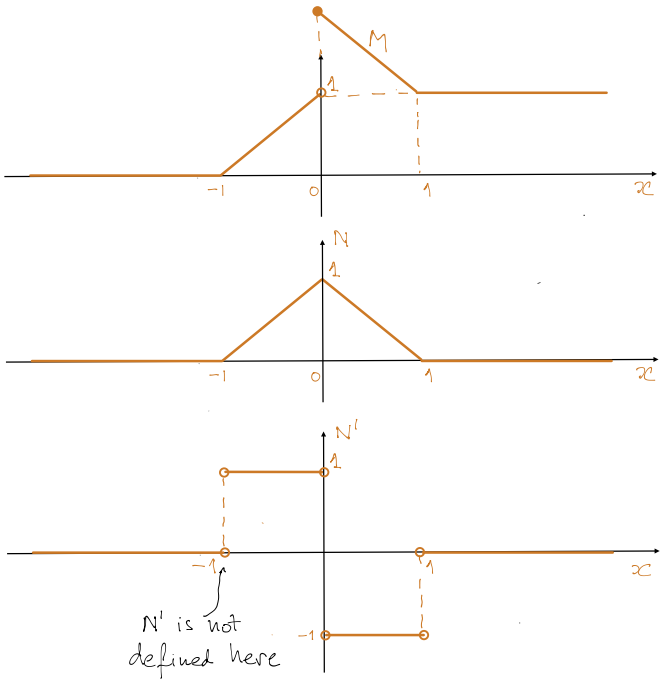


Figure 1.2 A hat function (middle), another function (top), and the common classical derivative (bottom).

and it is not defined at $x \in \{-1,0,1\}$. So, the domain of $N'(x)$ and $M'(x)$ is $\mathbb{R} \setminus \{-1,0,1\}$.

Consider then the following question. Let v be a smooth function over $\mathbb{R}$, such as $v(x) = \sin x$, how do we apply the integration-by-parts formula to the following integrals?

$$\int_{-1}^{1} N(x)v'(x)\,dx, \qquad \int_{-1}^{1} M(x)v'(x)\,dx. \tag{1.42}$$

Neither N nor M are smooth over the interval $(-1,1)$, so we need to proceed with caution. Notice, however, that both M and N are smooth over the intervals $(-1,0)$ and $(0,1)$, so we can proceed as follows, using u for either N or M,

$$\begin{aligned}
\int_{-1}^{1} u(x)v'(x)\,dx &= \int_{-1}^{0} u(x)v'(x)\,dx + \int_{0}^{1} u(x)v'(x)\,dx \\
&= \lim_{x\to 0^-} u(x)v(x) - u(-1)v(-1) - \int_{-1}^{0} u'(x)v(x)\,dx \\
&\quad + u(1)v(1) - \lim_{x\to 0^+} u(x)v(x) - \int_{0}^{1} u'(x)v(x)\,dx \\
&= u(1)v(1) - u(-1)v(-1) - \int_{-1}^{1} u'(x)v(x)\,dx \\
&\quad + v(0)\left(\lim_{x\to 0^-} u(x) - \lim_{x\to 0^+} u(x)\right)
\end{aligned} \tag{1.43}$$

The value

$$[\![u]\!]_{x=c} = \lim_{x\to c^-} u(x) - \lim_{x\to c^+} u(x) \tag{1.44}$$

is called the **jump** of u at $x = c \in \mathbb{R}$. Its value is equal to the jump discontinuity of u at $x = c$, so it is zero when u is continuous at $x = c$, and different than zero otherwise. For example, $[\![M]\!]_{x=0} = -1$. Hence, (1.43) for N and M is

$$\begin{aligned}
\int_{-1}^{1} N(x)\,v'(x)\,dx &= -\int_{-1}^{1} N'(x)\,v(x)\,dx \\
&= -\int_{-1}^{0} v(x)\,dx + \int_{0}^{1} v(x)\,dx \\
\int_{-1}^{1} M(x)\,v'(x)\,dx &= v(1)M(1) + v(0)[\![M]\!]_{x=0} - \int_{-1}^{1} M'(x)\,v(x)\,dx \\
&= v(1) - v(0) - \int_{-1}^{0} v(x)\,dx + \int_{0}^{1} v(x)\,dx.
\end{aligned}$$

So, the integration-by-parts formula applies as we know it for N over the interval $(-1,1)$, but not to M because of the discontinuity it has at $x = 0$. We generalize this observation next.

Functions N and M are called piecewise smooth, since each one of them has derivatives of any order in the open intervals $(-1,0)$ and $(0,1)$, but not on the entire real line. More generally, in the context of these notes, we say that a function $f\colon (a,b) \subset \mathbb{R} \to \mathbb{R}$ is **piecewise smooth** over (a,b) if there are $k \in \mathbb{N}$ points $a = c_0 \le \ldots \le c_k = b$ such that f is smooth in each interval (c_i, c_{i+1}) for $i = 0, k-1$.

☞ If both functions u and v are continuous and piecewise smooth, then the integration-by-parts formula used for smooth functions holds.

Theorem 1.2 (Integration by Parts Formula for Piecewise Smooth Functions)**.** *Let $(a,b) \subset \mathbb{R}$, and u, v be piecewise smooth functions. Let $c_0 = a \le \ldots \le c_k = b$ for $k \in \mathbb{N}$ be such that both u and v are smooth in each interval (c_i, c_{i+1}) for $i = 0, \ldots, k-1$. Then,*

$$\int_a^b u'(x)\,v(x)\,dx = \sum_{i=0}^{k} [\![u(x)\,v(x)]\!]_{x=c_i} - \int_a^b u(x)\,v'(x)\,dx, \tag{1.45}$$

where

$$\begin{aligned}
[\![u(x)\,v(x)]\!]_{x=c_0} &= -\lim_{x\to a^+} u(x)\,v(x) \\
[\![u(x)\,v(x)]\!]_{x=c_k} &= \lim_{x\to b^-} u(x)\,v(x).
\end{aligned}$$

The proof of this theorem is simple, and it follows the ideas we used in (1.43). It consists of decomposing the integral over (a,b) into a sum of integrals over (c_i, c_{i+1}) for $i = 0, \ldots, k-1$, and then integrating by parts in each one of these intervals, in which the two functions are smooth.

It follows from Thm. 1.2 that *if both u and v are continuous in (a,b) and piecewise smooth, then the same integration-by-parts used for smooth functions holds.*

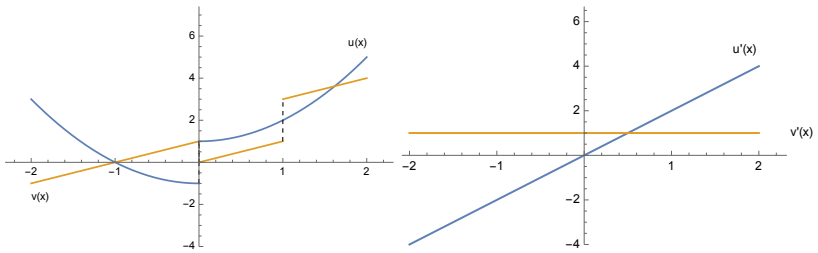
Example 1.24 Consider the functions

$$u(x) = \begin{cases} x^2 - 1 & x < 0, \\ x^2 + 1 & x \geq 0, \end{cases} \quad \text{and} \quad v(x) = \begin{cases} x + 1 & x < 0, \\ x & x \in [0,1], \\ x + 2 & x > 1. \end{cases}$$

Their derivatives are

$$u'(x) = 2x \text{ for } x \neq 0, \quad \text{and} \quad v'(x) = 1 \text{ for } x \notin \{0,1\}.$$

The two functions and their derivatives are plotted below:



Consider the following expression

$$\int_{-2}^{2} u(x)\, v'(x)\, dx. \tag{1.46}$$

Its value can be readily computed by direct integration, and it is

$$\int_{-2}^{2} u(x)\, v'(x)\, dx = \int_{-2}^{0} x^2 - 1\, dx + \int_{-2}^{0} x^2 + 1\, dx = \frac{16}{3}.$$

Let's integrate (1.46) by parts, and verify that returns the same value. To this end, notice that u is smooth in the intervals $(-2,0)$ and $(0,2)$, while v is smooth in the intervals $(-2,0)$, $(0,1)$, and $(1,2)$. The two functions are piecewise smooth in $(-2,2)$ if we select $k = 3$ and $c_0 = -2, c_1 = 0, c_2 = 1, c_3 = 2$. Notice that u is smooth in $(0,2)$, and hence it is automatically smooth in the two intervals (c_1, c_2) and (c_2, c_3). From (1.45),

$$\begin{aligned}
\int_{-2}^{2} u(x)\, v'(x)\, dx &= [\![u(x)\, v(x)]\!]_{x=-2} + [\![u(x)\, v(x)]\!]_{x=0} + [\![u(x)\, v(x)]\!]_{x=1} \\
&\quad + [\![u(x)\, v(x)]\!]_{x=2} - \int_{-2}^{2} u'(x)\, v(x)\, dx \\
&= -u(-2)\, v(-2) + \lim_{x \to 0^-} u(x)\, v(x) - \lim_{x \to 0^+} u(x)\, v(x) \\
&\quad + \lim_{x \to 1^-} u(x)\, v(x) - \lim_{x \to 1^+} u(x)\, v(x) + u(2)\, v(2) \\
&\quad - \int_{-2}^{0} 2x(x+1)\, dx - \int_{0}^{1} 2x.x\, dx - \int_{1}^{2} 2x(x+2)\, dx \\
&= -3.(-1) + (-1).1 - 1.0 + 2.1 - 2.3 + 5.4 - \frac{38}{3} \\
&= \frac{16}{3},
\end{aligned}$$

and we verified the identity.

An alternative way to obtain the same result is to split the integral as a sum of integrals over $(-2,0)$, $(0,1)$ and $(1,2)$, and then integrate each one of these integrals by parts. By rearranging the terms, we will arrive to the expression that we used from Thm. 1.2.

1.2 Linear Algebra for Spaces of Functions

The formulation of finite element methods is more easily performed and understood with some basic concepts of linear algebra, in this case, applied to spaces of functions.

1.2.1 Vector Spaces of Functions

The first encounter with finite element methods is for many the first encounter with the use of vector spaces in a context other than one in which vectors represent elementary physics quantities, such as forces or velocities. It is also the first encounter with infinite-dimensional vector spaces. In studying finite element methods, we are interested in vector spaces in which each vector is a function, that is, in *vector spaces of functions*.

For example, consider the set V of all real quadratic polynomials that are zero at zero, namely, functions of the form

$$f(x) = ax^2 + bx$$

for any $a, b \in \mathbb{R}$, such as $f_1(x) = 3x^2$ and $f_2(x) = x$. Notice that $f_1(x) + f_2(x) = 3x^2 + x$ is also a function in V, and $3f_2(x) = 3x$ is another one, so addition of two polynomials in V returns a polynomial in V, and multiplication of a polynomial in V by a scalar (real number) is also a polynomial in V.

We can think of each polynomial in V as the vector in $\mathbb{R}^2$ that starts at the origin and ends at the point with coordinates (a,b), see Fig. 1.3. The sum of two functions in V corresponds to adding the two vectors, and similarly with the multiplication by a scalar (real number). You may be wondering about why to call each function a vector, or simply why to talk about vector spaces of functions? With this identification we can define the *dimension* and a basis for V, and by using a basis, we will be able to build any function (vector) in the space.

We begin by reviewing the definition of vector spaces.

Definition 1.1 (Vector Space)**.** *A Vector Space V is a set for which two operations $+$ and $\cdot$ are defined, called* **vector addition** *and* **multiplication by a scalar**, *such that for all $u, v, w \in V$ and all $\alpha, \beta \in \mathbb{R}$ they satisfy:*

1. **Closure**: *$u + v \in V$, and $\alpha \cdot u \in V$.*
2. **Commutativity:** *$u + v = v + u$.*

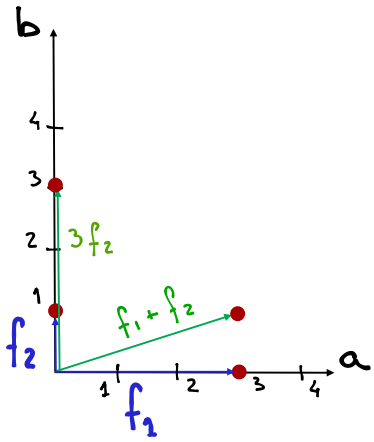


Figure 1.3 Identification of polynomials with vectors in $\mathbb{R}^2$.

3. **Associativity:** $u+(v+w)=(u+v)+w$, *and* $\alpha\cdot(\beta\cdot u)=(\alpha\beta)\cdot u$.

4. **Identity:** *There exists an element* $0\in V$, *called "zero," such that* $u+0=u$, *and* $1\cdot u=u$.

5. **Additive Inverse:** *For any* $u\in V$, *there exists* $v\in V$ *such that* $v+u=0$.

6. **Distributivity:** $(\alpha+\beta)\cdot u=\alpha\cdot u+\beta\cdot u$ *and* $\alpha\cdot(u+v)=\alpha\cdot u+\alpha\cdot v$.

The elements of V *are called vectors.*

In the following, we will drop the symbol $\cdot$ to indicate multiplication by a scalar, unless there is ambiguity. So, for example, for $\alpha\cdot u$ will write $\alpha\, u$.

You are by now very familiar with $\mathbb{R}^n$ as a vector space, $n\in\mathbb{N}$. Each point of $\mathbb{R}^n$ defines a vector under the standard definition of vector addition and multiplication by a scalar in $\mathbb{R}^n$. As aforementioned, what might be new for you is that sets of functions can also be vector spaces. In this case, each function in the set is a "vector."

To complete the depiction of functions as elements of a vector space, we need to specify the **vector-addition** and **multiplication-by-a-scalar** operations. Fortunately, they are defined exactly as you would expect: Let V be a set of functions over a domain $\Omega\subseteq\mathbb{R}^n$, and let $f,g\in V$ and $\alpha\in\mathbb{R}$, then

- **Vector addition:** the function defined as $h(x)=f(x)+g(x)$ for all $x\in\Omega$.
- **Multiplication by a scalar:** the function defined as $w(x)=\alpha f(x)$ for all $x\in\Omega$.

To illustrate this definition, we will consider sets of *smooth functions*, which as in previous sections, are functions in which all derivatives exist and are continuous.

Examples:

1.25 The set $V_1=\{f\colon[a,b]\to\mathbb{R}\text{ smooth}\}$ is a vector space. We can check each one of the non-trivial properties:

i. Closure: If $u,v\in V_1$ and $\alpha\in\mathbb{R}$, then $u+v\in V_1$ and $\alpha u\in V_1$, since the sum of smooth functions is another smooth function, and multiplication of a smooth function by a scalar is anoter smooth function.

ii. Identity: the function $z(x)=0$ for all $x\in[a,b]$ is the "zero" of the space.

The rest of the properties are easy to check.

1.26 The set $V_2=\{f\colon[a,b]\to\mathbb{R}\text{ smooth }\mid w(a)=w(b)=0\}$ is a vector space. Since $V_2\subset V_1$ in Example 1.25, V_2 inherits commutativity, associativity, and distributivity in Def. 1.1 from V_1. We need to

only check closure, identity and additive inverse. Closure follows because if $u, v \in V_2$, and $w = u + v$, then $w(a) = u(a) + v(a) = 0$ and $w(b) = u(b) + v(b) = 0$, and hence $w \in V_2$. Since the zero function is in V_2, identity is satisfied. Finally, if $u \in V_2$, then $-u \in V_2$ because $u(a) = -u(a) = 0$ and $u(b) = -u(b) = 0$, and hence additive inverse is also satisfied.

1.27 The set of polynomials of degree less or equal than $k \in \mathbb{N} \cup \{0\}$ over an interval $I \subset \mathbb{R}$, $\mathbb{P}_k(I)$, is a vector space. This is because the sum of polynomials in $\mathbb{P}_k$ and the product by a scalar is still a polynomial of degree less or equal than k (closure), and because the function $0 \in \mathbb{P}_k$ (identity). The rest of the properties are easy to check.

In addition to vector spaces, we will use a closely related concept, that of an affine subspace, which we define next.

Definition 1.2 (Vector Subspace)**.** *Let W be a vector space. A vector space $V \subset W$ is vector subspace of W.*

Definition 1.3 (Affine Subspace)**.** *Let W be a vector space. An affine subspace of W is a set $S \subset W$ such that for any $s_1 \in S$ the set*

$$V = \{s_2 - s_1 \mid s_2 \in S\}$$

is a vector subspace of W. The vector space V is called the **direction** *of S, or the associated vector space to S.*

The direction V is independent of the choice of s_1.

The direction V is independent of the choice of s_1.

To see this, let $s_a, s_b \in S$ and

$$V_a = \{s - s_a \mid s \in S\}, \qquad V_b = \{s - s_b \mid s \in S\}$$

be the associated vector spaces to S. We will prove that $V_a = V_b$, and hence that the direction is independent of s_1. We will use the fact that $\Delta s = s_a - s_b$ belongs to both V_a and V_b, by definition. So, if $v \in V_a$, then there exists $\overline{s} \in S$ such that

$$\overline{s} = v + s_a = v + \Delta s + s_b,$$

and hence $w = v + \Delta s = \overline{s} - s_b \in V_b$. But $w = v + \Delta s \in V_a$, since $\Delta s \in V_a$, from where $v = w - \Delta s \in V_b$ as well, since $\Delta s \in V_b$. We conclude then that if $v \in V_a$, $v \in V_b$, or $V_a \subseteq V_b$. A similar argument leads to $V_a \supseteq V_b$, and hence to $V_a = V_b$.

Of course, using the notation from the definition, a vector space $V \subset W$ is an affine subspace of W. Elements of an affine subspace are called points and not vectors, since it is not a vector space.

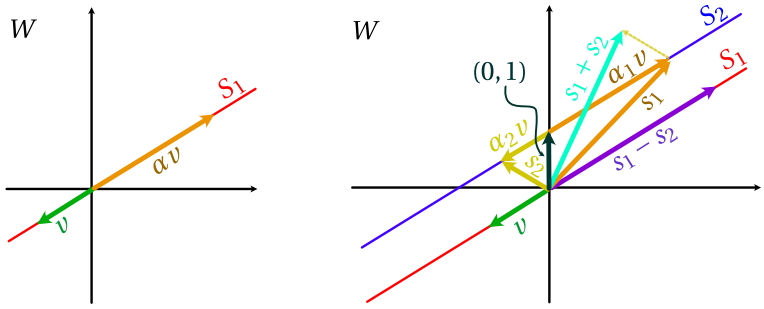


Figure 1.4 Sketch of sets S_1 and S_2 in Example 1.28. The former is a vector space, while the latter is not; it is an affine subspace of W instead.

Examples:

1.28 Let $v = (-1,-1) \in W = \mathbb{R}^2$. The set $S_1 = \{\alpha v \mid \alpha \in \mathbb{R}\}$ is a vector space. Instead, the set $S_2 = \{\alpha v + (0,1) \mid \alpha \in \mathbb{R}\}$ is *not* a vector space. This is because if $s_1 = \alpha_1 v + (0,1)$ and $s_2 = \alpha_2 v + (0,1)$, then $h = s_1 + s_2 = \alpha_1 v + (0,1) + \alpha_2 v + (0,1) = (\alpha_1 + \alpha_2) v + (0,2)$, and hence $h \notin S_2$.
Instead, the set S_2 is an affine subspace of W, since $s_1 - s_2 = (\alpha_1 - \alpha_2) v$, and hence $s_1 - s_2$ can be any element of the vector space S_1. Please see Fig. 1.4 for a sketch.

1.29 The set $V_3 = \{w\colon [a,b] \to \mathbb{R} \text{ smooth } \mid w(a) = 1 = w(b) = 1\}$ is *not* a vector space. This is because if $u, v \in V_3$, and $w = u + v$, then $w(a) = u(a) + v(a) = 2$, and similarly for $w(b)$, and hence $w \notin V_3$.
Instead, V_3 is an affine subspace of the vector space $W = \{w\colon [a,b] \to \mathbb{R} \text{ smooth}\}$. To see this, first notice first that for any $v_1, v_2 \in V_3$, $v_2 - v_1 \in V_2$ of Example 1.26, a vector space, since $v_1(a) - v_2(a) = 0 = v_1(b) - v_2(b)$. This implies that for any $v_1 \in V_3$,

$$V = \{v_2 - v_1 \mid v_2 \in V_3\} \subseteq V_2.$$

Second, notice that if $v \in V_2$, then $v_1 + v = v_2 \in V_3$, so $v_2 - v_1 = v$, from where

$$V \supseteq V_2.$$

Therefore, $V = V_2$, and hence V_3 is an affine subspace of W with V_2 as its direction.

Vector spaces that are subspaces of $\mathbb{R}^n$ can be identified with (hyper-)planes that contain the origin. Affine subspaces of $\mathbb{R}^n$ are (hyper-)planes that may not contain the origin, and hence, will be parallel to a vector subspace, their direction, as illustrated by S_1 and S_2 in Fig. 1.4. Also in the figure, notice that the role of

the vector $(0,1)$ is to "transport" the vector space S_1 parallel to itself to become the affine subspace S_2. More generally, if s_1 is any element of an affine subspace S, then any other element $s \in S$ can be written as $s = s_1 + w$ for $w \in V$, since by the definition of an affine subspace, $s - s_1 = w \in V$. In Fig. 1.4, when $s = s_2$, then $w = \alpha_2 v$.

More importantly, if $s_1 \in S$, then we can reach all elements in S by adding an element in its direction V, namely,

$$S = \{s_1 + w \mid w \in V\}. \tag{1.47}$$

$S = \{s_1 + w \mid w \in V\}$ **for any** $s_1 \in S$**.**

To see this, let

$$U = \{s_1 + w \mid w \in V\}.$$

We want to show that $\mathscr{U} = S$. For any $s_2 \in S$, $w = s_2 - s_1 \in V$, by definition, so $s_2 = w + s_1 \in U$. Therefore $U \supseteq S$.

To see that $U \subseteq S$, notice that for any $w \in V$, there exists $s_2 \in S$ such that $s_2 - s_1 = w$, since all elements of V are generated by such differences. This implies that $s_2 = s_1 + w$, and hence that $w + s_1 \in S$, or $U \subseteq S$.

1.2.1.1 Bases in a vector space of functions

Next, we review the definition of a basis for a vector space, and see examples of bases in vector spaces of functions.

Definition 1.4 (Linear combinations, or span)**.** *Let V be a vector space and $U \subset V$ be a set of vectors in V. The* **span** *of U, span(U), is the set*

$$span(U) = \left\{ \sum_{i=1}^{n} c_i e_i \mid n \in \mathbb{N}, e_i \in U, c_i \in \mathbb{R} \right\}. \tag{1.48}$$

The set span(U) contains all **linear combinations** *of vectors in the set U.*

In the following, we denote vectors in $\mathbb{R}^n$ by the Cartesian coordinates of their end points.

Examples:

1.30 $U_1 = \{e_1, e_2\} \subset \mathbb{R}^3$, where $e_1 = (1,0,0)$ and $e_2 = (1,0,1)$. Then

$$\begin{aligned} \text{span}(U_1) &= \{c_1 e_1 + c_2 e_2 \mid (c_1, c_2) \in \mathbb{R}^2\} \\ &= \{(c_1 + c_2, 0, c_2) \in \mathbb{R}^3 \mid (c_1, c_2) \in \mathbb{R}^2\} \end{aligned}$$

is the plane that contains 0 and vectors e_1 and e_2, or whose normal is in the direction $e_1 \times e_2$, where $'\times'$ is the vector cross-product. Notice that vectors in U do not need to be unit vectors.

1.31 $U_2 = \{1, x, x^2\}$. Then span(U_2) is the set of all quadratic polynomials. The vector with components $(3,4,5)$, or $c_1 = 3$, $c_2 = 4$ and $c_3 = 5$ is the polynomial

$$p(x) = 3 + 4x + 5x^2,$$

and the quadratic polynomial $q(x) = 5 - 2x - 6x^2$ has components $(5,-2,-6)$ in this basis.

1.32 $U_3 = \{e_1, \ldots, e_k\} \subset V$. Then

$$\text{span}(U_3) = \sum_{i=1}^{k} c_i e_k = c_1 e_1 + \ldots + c_k e_k.$$

This examples illustrates the construction of the span of a set made of a finite number of vectors in V.

1.33 Let $W = \{(0, y, z) \mid (y, z) \in \mathbb{R}^2\}$ and $U_4 = \{(1,0,0)\} \cup W$. Then, span(U_4) = $\mathbb{R}^3$. For example, if $w = (2,3,4) \in \mathbb{R}^3$, we can write it as $w = 2 \times (1,0,0) + 1 \times (0,3,4)$. In Def. 1.4, this follows after setting $n = 2$, $c_1 = 2$, $c_2 = 1$, $e_1 = (1,0,0)$ and $e_2 = (0,3,4)$. Alternatively, we could have written $w = 2 \times (1,0,0) + 2 \times (0,3/2) + 8 \times (0,0,1/2)$, in which we set $n = 3$, $c_2 = 2$, $c_3 = 8$, $e_2 = (0,3/2,2)$ and $e_3 = (0,0,1/2)$ instead.

When U has an infinite number of vectors, elements of span(U) are formed through linear combinations of any finite number of vectors in U.

Definition 1.5 (Linearly Independent and Linearly Dependent Set of Vectors)**.** *Let V be a vector space and let $e_i \in V$ for $i = 1, \ldots, n$. The set of vectors $U = \{e_1, \ldots, e_n\}$ is linearly independent whenever*

$$\sum_{i=1}^{n} c_i e_i = 0 \Longleftrightarrow c_i = 0 \text{ for } i = 1, \ldots, n. \tag{1.49}$$

Otherwise, the set of vectors U is linearly dependent.

Examples:

1.34 $U_1 = \{e_1, e_2\} \subset \mathbb{R}^3$, where $e_1 = (1,0,0)$ and $e_2 = (1,0,1)$. The set of vectors in U_1 is linearly independent:

$$c_1(1,0,0) + c_2(1,0,1) = (0,0,0) \Longleftrightarrow \begin{cases} c_1 + c_2 = 0 \\ c_2 = 0 \end{cases} \Longleftrightarrow c_1 = c_2 = 0.$$

1.35 $U_2 = \{1, x, x^2\}$. The set of vectors U_2 is linearly independent:

$$p(x) = c_1 + c_2 x + c_3 x^2 = 0 \quad x \in [0,1] \Longleftrightarrow c_1 = c_2 = c_3 = 0.$$

To see this, it is enough to evaluate $p(x)$ at three different locations, for example. Say, $p(0) = 0$, $p(1/2) = 0$, $p(1) = 0$. The resulting system of equations has $c_1 = c_2 = c_3 = 0$ as the only solution.

1.36 $U_3 = \{1, x, 2+3x\}$. This is a linearly dependent set of vectors, since if we let $e_1 = 1$, $e_2 = x$, and $e_3 = 2+3x$, then $2e_1 + 3e_2 - e_3 = 0$.

1.37 Consider the set $U_4 = \{\min\{0, x\}, x\}$ of functions with domain $[a, b]$. If $[a, b] = [-1, 1]$, U_4 is a set of linearly independent functions, since for

$$f(x) = c_1 \min\{0, x\} + c_2 x,$$

we have that

$$0 = f(-1) = -c_2, \qquad 0 = f(1) = c_1 + c_2 \Longrightarrow c_1 = c_2 = 0.$$

Instead, if $[a, b] = [0, 1]$, this is a linearly dependent set. To see this, notice that for $x \in [0, 1]$, $\min\{0, x\} = x$, so the two functions are precisely the same function over this interval.

Definition 1.6 (Basis and Dimension of a Vector Space)**.** *Let V be a vector space and $e_i \in V$ for $i = 1, \ldots, n$, $n \in \mathbb{N}$. The set $U = \{e_1, \ldots, e_n\}$ is a basis of V if U is linearly independent and span$(U) = V$. The number of vectors in a basis is the dimension of V.*

A vector space that does not have a basis is an infinite-dimensional space.

Given a basis $U = \{e_1, \ldots, e_n\}$ in a vector space V, and a vector $v \in V$, there there exists a *unique* set of numbers $(c_1, \ldots, c_n) \in \mathbb{R}^n$ such that

$$v = c_1 e_1 + \ldots + c_n e_n. \tag{1.50}$$

The numbers $c_1, \ldots, c_n$ are called the **components** of v in basis U.

Conversely, when the components $(c_1, \ldots, c_n)$ span all points in $\mathbb{R}^n$, the vector v in (1.50) spans the space V. Because all possible vectors in V are obtained by evaluating all possible values of $(c_1, \ldots, c_n)$, the variables $c_1, \ldots, c_n$ are called **degrees of freedom** of V.

Examples:

1.38 Consider the set $U_1 = \{e_1, e_2\} \subset \mathbb{R}^3$, where $e_1 = (1, 0, 0)$ and $e_2 = (1, 0, 1)$. The set of vectors in U_1 is not a basis for $\mathbb{R}^3$, since the vector $(0, 1, 0) \notin$ span(U_1).

1.39 Consider the set $U_4 = \{e_1, e_2, e_3\} \subset \mathbb{R}^3$, where $e_1 = (1, 0, 0)$, $e_2 = (1, 0, 1)$, and $e_3 = (0, 1, 0)$. The set U_4 is a basis for $\mathbb{R}^3$, since it can be seen to be linearly independent, and any vector in $\mathbb{R}^3$ is a linear combination of the basis: If $(x, y, z) \in \mathbb{R}^3$, then $(x, y, z) = (x - y)e_1 + ye_2 + ze_3$. The dimension of $\mathbb{R}^3$ is then 3.

1.40 The set $U_2 = \{1, x, x^2\} \subset V_1 = \{f : (0, 1) \to \mathbb{R} \text{ smooth}\}$. The set of vectors U_2 is not a basis for V_1. In fact, there is no basis for V_1, and hence it is an **infinite dimensional space**.

The set U_2 is a basis for the vector space $\mathbb{P}_2$ formed by all quadratic polynomials, whose dimension is 3.

1.41 The set $U_5 = \{1+x, x-x^2, x^2-1\}$ is another basis for $\mathbb{P}_2$. To see this, notice that given a polynomial $p(x) = a+bx+cx^2 \in \mathbb{P}_2$ for $a,b,c \in \mathbb{R}$, we can write it as

$$p(x) = \frac{a+b+c}{2}(1+x) + \frac{b-c-a}{2}(x-x^2) + \frac{b+c-a}{2}(x^2-1).$$

Instead, the set $U_5' = \{1+x, x-x^2, x^2+1\}$ is not a basis, since the three functions are not linearly independent: $(x^2+1)+(x-x^2) = 1+x$.

1.42 The set $U_6 = \{\sin(x), \sin(2x), \sin(3x), \sin(4x)\}$ is a basis for $\text{span}(U_6)$ over the interval $(0, 2\pi)$, since the 4 functions are linearly independent. One way to see the linear independence is as follows: If we have

$$\sum_{i=0}^{n} c_i \sin(ix) = 0$$

with $n = 4$ here, we need to show that this implies that $c_i = 0$ for any i. This follows by multiplying the last equation by $\sin(jx)$ for any $j \in \{1, \dots, n\}$ and integrating over the interval $(0, 2\pi)$. In this case we get that

$$\sum_{i=0}^{n} c_i \int_0^{2\pi} \sin(ix)\sin(jx)\,dx = 0. \tag{1.51}$$

We then notice that

$$\int_0^{2\pi} \sin(ix)\sin(jx)\,dx = \begin{cases} \pi & i = j, \\ 0 & i \neq j. \end{cases}$$

Using this in (1.51) allows us to conclude that

$$0 = \sum_{i=0}^{n} c_i \underbrace{\int_0^{2\pi} \sin(ix)\sin(jx)\,dx}_{\neq 0 \text{ only if } i=j} \Longrightarrow c_j = 0,$$

and since this is true for any j, we can conclude that U_6 is a linearly independent set, and hence it is a basis for $\text{span}(U_6)$.

1.2.1.2 Linear functional and bilinear form

We conclude this section by introducing two more definitions, which will allow us to talk about variational equations in an abstract way.

Definition 1.7 (Linear Functional)**.** *Let V be a vector space. A linear functional is a function $\ell : V \to \mathbb{R}$ such that for any $u, v \in V$ and $\alpha \in \mathbb{R}$*

$$\ell(u + \alpha v) = \ell(u) + \alpha \ell(v). \tag{1.52}$$

Examples:

1.43 Let $V_1 = \{f : [0,1] \to \mathbb{R} \text{ smooth}\}$, then

$$\ell(v) = \int_0^1 x^2 v(x)\, dx$$

is a linear functional in V_1. This is because:

- The value of $\ell(v)$ can be computed for any function $v \in V_1$, so it is defined for *any* function in V_1.
- It is simple to see that (1.52) is true, to wit, for $u, v \in V_1$ and $\alpha \in \mathbb{R}$,

$$\begin{aligned}\ell(u+\alpha v) &= \int_0^1 x^2(u(x)+\alpha v(x))\, dx \\ &= \int_0^1 x^2 u(x)\, dx + \alpha \int_0^1 x^2 v(x)\, dx \\ &= \ell(u) + \alpha \ell(v).\end{aligned}$$

Let's compute the value of the linear functional for a couple of functions:

- $\ell(\cos(x)) = \int_0^1 x^2 \cos(x)\, dx = 2\cos(1) - \sin(1)$.
- $\ell(x^4) = \int_0^1 x^2 x^4\, dx = \frac{x^7}{7}\Big|_{x=1} = \frac{1}{7}$.

This is an example of linear functionals of the form

$$\ell(v) = \int_a^b f(x) v(x)\, dx$$

for some function f, which we will encounter often in later sections.

1.44 Let $V = \mathbb{R}^2$, and $f = (f_1, f_2) \in \mathbb{R}^2$. For $v = (v_1, v_2) \in V$,

$$\ell(v) = v_1 f_1 + v_2 f_2 \tag{1.53}$$

is a linear functional.

1.45 Let V be the set of continuous functions over $\mathbb{R}$. For $v \in V$, let

$$\ell(v) = v(0) \tag{1.54}$$

This is a linear functional. You may have encountered this functional written in a different way:

$$\ell(v) = \int_{\mathbb{R}} \delta(x) v(x)\, dx,$$

namely, using the *Dirac delta function*. A problem with the denomination of $\delta(x)$ as a function is that $\delta(x)$ is not a function over the real line, since it does not assign a value to points in the real line. Instead, it is a linear functional, since it assigns a scalar value to each function over the real line.

A linear functional is also called a **one-form**.

Definition 1.8 (Bilinear Form)**.** *Let V be a vector space. A bilinear form is a function $a\colon V \times V \to \mathbb{R}$ that is linear in each argument. More precisely, for any $u, v, w \in V$ and $\alpha \in \mathbb{R}$*

$$\begin{aligned} a(u+\alpha v, w) &= a(u,w) + \alpha a(v,w) \\ a(w, u+\alpha v) &= a(w,u) + \alpha a(w,v). \end{aligned} \tag{1.55}$$

If, additionally, for all $u, v \in V$

$$a(u,v) = a(v,u), \tag{1.56}$$

then "a" is a symmetric bilinear form.

Examples:

1.46 Let $V_1 = \{f\colon [0,1] \to \mathbb{R} \text{ smooth}\}$, then

$$a(u,v) = \int_0^1 u'(x)\,v'(x)\,dx$$

is a bilinear form, since $a(u,v)$ can be computed for any functions $u, v \in V_1$, and it is simple to see that (1.55) is true. To wit,

$$\begin{aligned} a(u+\alpha v, w) &= \int_0^1 (u' + \alpha v')\,w'\,dx \\ &= \int_0^1 u'w'\,dx + \alpha \int_0^1 v'w'\,dx \\ &= a(u,w) + \alpha a(v,w), \end{aligned}$$

and similarly with the other slot.

This bilinear form is symmetric.

Let's compute the value of the bilinear form for a few functions:

- $a(\sin(x), x^2) = \int_0^1 \cos(x) 2x\,dx = 2(\sin(1) - \cos(1))$.
- $a((x-1)^2, x^3) = \int_0^1 2(x-1)3x^2\,dx = -1/2$.

1.47 Let $V = \mathbb{R}^2$, and

$$M = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix},$$

where $u = (u_1, u_2)$ and $v = (v_1, v_2)$ are vectors in $\mathbb{R}^2$, written in a column matrix form. Then, we can define

$$\begin{aligned} a(u,v) &= u^{\mathsf{T}} M v \\ &= \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ &= 2u_1v_1 + 3u_2v_1 + u_1v_2 + 2u_2v_2. \end{aligned} \tag{1.57}$$

This is a bilinear form in V. It is *not* a symmetric bilinear form. For example, $a((0,1),(2,0)) = 6$, and $a((2,0),(0,1)) = 2$.

Now, if

$$M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{1.58}$$

then $a(u,v) = u \cdot v$, or the dot product between vectors u and v. So, the dot product *is* a bilinear form.

1.48 Let V_1 be that of Example 1.46, and set

$$a(u,v) = \int_0^1 \sin(x)u(x)v'(x)\,dx + u'(1/2)v(1/2).$$

This is a bilinear form, since it is defined for any pair of smooth functions, and it is linear in each argument. It is not symmetric, as it can be inferred from the different roles u and v play in each term. An alternative way to see this is by choosing two functions u and v and evaluating $a(u,v)$; there is a high chance that $a(u,v)$ will be different than $a(v,u)$ if a is not symmetric, and single pair of functions for which $a(u,v) \neq a(v,u)$ is enough to show that it is not symmetric. Set $u(x) = x^2$ and $v(x) = x^3$, then

$$a(x^2,x^3) = \int_0^1 \sin(x)x^2 3x^2\,dx + 2(1/2)(1/2)^3 = 577/8 - 39\cos(1) - 60\sin(1),$$
$$a(x^3,x^2) = \int_0^1 \sin(x)x^3 2x\,dx + 3(1/2)^2(1/2)^2 = 771/16 - 26\cos(1) - 40\sin(1).$$

so $a(x^2,x^3) \neq a(x^3,x^2)$, and this proves that a is not symmetric.

1.2.2 Linear Variational Equations

Having defined linear functionals and bilinear forms, we can now write the variational equations we have seen so far in a simple, abstract way. In paarticular, we are now ready to give a proper definition of a variational equation.

Definition 1.9 (Variational Equation)**.** *Let $\mathcal{W}, \mathcal{V}$ be vector spaces and $F\colon \mathcal{W} \times \mathcal{V} \to \mathbb{R}$ be linear in the second argument, that is,*

$$F(u, v + \alpha w) = F(u,v) + \alpha F(u,w)$$

for any $u \in \mathcal{W}$, $v, w \in \mathcal{V}$ and $\alpha \in \mathbb{R}$.

A variational equation is an equation of the form

$$F(u,v) = 0 \qquad \forall v \in \mathcal{V}, \tag{1.59}$$

The space $\mathcal{V}$ is called the test space.

Definition 1.10 (Linear Variational Equation)**.** *A linear variational equation is a variational equation that is linear in the first argument.*

For completeness, F from (1.59) is linear in the first argument if

$$F(u+\alpha w, v) = F(u,v)+\alpha F(w,v)$$

for any $u, w \in \mathcal{W}$, $v \in \mathcal{V}$ and $\alpha \in \mathbb{R}$.

The variational equations we have seen in §1.1.2 are all linear variational equations that have the abstract form

$$0 = F(u,v) = a(u,v) - \ell(v) \qquad \forall v \in \mathcal{V}, \tag{1.60}$$

or

$$\boxed{a(u,v) = \ell(v) \qquad \forall v \in \mathcal{V}} \tag{1.61}$$

where $a\colon \mathcal{W}\times\mathcal{W} \to \mathbb{R}$ is a bilinear form, and $\ell\colon \mathcal{W} \to \mathbb{R}$ is a linear functional.

Examples:

1.49 Let's identify the bilinear form and linear functional for the model problem. Consider the variational equation (1.9a) that the solution u of Problem 1.1 satisfies, namely,

$$\int_\Omega \left[k(x)u'(x)v'(x) + b(x)u(x)'v(x) + c(x)u(x)v(x)\right]\,dx - k(L)d_L v(L) = \int_\Omega f(x)v(x)\,dx. \tag{1.62}$$

for any $v \in \mathcal{V} = \{w\colon \Omega \to \mathbb{R} \text{ smooth } \mid w(0) = 0\}$, with $\Omega = (0,L)$.
Let's identify

$$a(u,v) = \int_\Omega \left[k(x)u'(x)v'(x) + b(x)u(x)'v(x) + c(x)u(x)v(x)\right]\,dx, \tag{1.63a}$$

$$\ell(v) = \int_\Omega f(x)v(x)\,dx - k(L)d_L v(L), \tag{1.63b}$$

as a bilinear form and a linear functional, respectively.

1.50 Consider the differential equation

$$u(x)u'(x) + u''(x) = 0$$

for all $x \in (0,L)$. This is a nonlinear equation because of the first term. If u satisfies it, then it also satisfies the variational equation

$$F(u,v) = \int_0^L u(x)u'(x)v(x) - u'(x)v'(x)\,dx = 0 \tag{1.64}$$

for any $v \in \mathcal{V} = \{w\colon [0,L] \text{ smooth } \mid v(0) = v(L) = 0\}$. This equation follows after multiplying the differential equation by v, integrating over $(0,L)$, and then integrating by parts the second term.
Equation (1.64) is a variational equation, but because F is non-linear in the first argument, it is not a linear variational equation.

The specification of the space $\mathcal{W}$ to build the domains of a and ℓ is not important, at least at this stage, so we will skip it.

1.3 Variational Numerical Methods

The abstract variational equation (1.61) inspires the formulation of a class of **numerical methods**. However, before describing them, we will discuss what a numerical method is.

What is a numerical method? A numerical method for a problem such as Problem 1.1 is a definition of a sequence of functions $\{u_{h1}, u_{h2}, \dots u_{hn}, \dots\}$ that *can be computed* and such that they *approximate* the exact solution u as $n \to \infty$.

The phrase "can be computed" means that functions $u_{h1}, u_{h2}, \dots$ are not implicitly defined but their values can be explicitly computed, limited only by available computational resources.

The most important requirement, however, is that the sequence $u_{h1}, u_{h2}, \dots u_{hn}, \dots$ approximates the solution of the problem to any degree of accuracy provided we choose n large enough. For example, the method could guarantee that for some value n the maximum difference between the u and u_{hn} is smaller than a desired tolerance everywhere in the domain.

In the following, we will be concerned with defining finite element methods and their implementation, and postpone the discussion of their approximation properties to §**??**. In other words, we will first discuss *what* finite element methods are, and then we will discuss *how* finite element methods provide approximations.

1.3.1 Variational Methods

A variational numerical numerical method defines approximations by using the following observation. Consider a solution u of a problem that satisfies a variational equation of the form

$$F(u, v) = 0 \qquad \forall\, v \in \mathcal{V}$$

for some function F and test space $\mathcal{V}$.

A variational method is defined by finite dimensional spaces function spaces $\mathcal{V}_h$ and $\mathcal{S}_h$. It defines an approximation u_h of u by finding $u_h \in \mathcal{S}_h$ that satisfies that

$$F(u_h, v_h) = 0 \qquad \forall\, v \in \mathcal{V}_h. \tag{1.65}$$

It is also possible to consider a variational equation that changes with h, i.e. a function F_h. We will see examples of this later.

The space $\mathcal{S}_h$ is called the **trial space**, it is an *affine space* where u_h is sought. The space $\mathcal{V}_h$ is a *vector space* that acts as the test space for variational equation (1.65). By prescribing a way to select $\mathcal{V}_h$, $\mathcal{S}_h$ for different values of h, the method can define a sequence $\{u_{h1}, u_{h2}, \dots\}$ that approximates u.

The simplest scenario, and the one we will be concerned with, is that in which F defines a linear variational equation and the test space is the direction of the trial space. Such methods define the approximation u_h as the solution of a problem of the following type:

Problem 1.2 (Variational Method)**.** *Let $\mathcal{W}_h$ be a finite-dimensional vector space, $a\colon \mathcal{W}_h \times \mathcal{W}_h \to \mathbb{R}$ be a bilinear form, $\ell\colon \mathcal{W}_h \to \mathbb{R}$ be a linear functional, $\mathcal{S}_h \subseteq \mathcal{W}_h$ and $\mathcal{V}_h \subseteq \mathcal{W}_h$, with $\mathcal{V}_h$ the direction of $\mathcal{S}_h$.*

$$\text{Find } u_h \in \mathcal{S}_h \text{ such that } a(u_h, v_h) = \ell(v_h) \text{ for all } v_h \in \mathcal{V}_h. \tag{1.66}$$

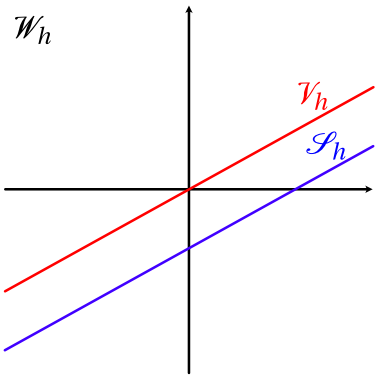


Figure 1.5 Illustration of the relationship among spaces for the variational method in Problem 1.2.

The variational equation F and the trial and test spaces need to be selected so that they can provide an approximation to the solution of the problem of interest.

Different choices of a, ℓ, and $\mathcal{S}_h$ lead to different variational methods.
Finite element methods are a type of variational numerical methods.

Let's look at a simple example of such problem and its solution.

Example 1.51 Consider the linear variational equation in Example 1.49 with $\Omega = [0,1]$, $k(x) = 1$, $b(x) = c(x) = 0$, $f(x) = 1$ and $d_L = 0$.

Let

$$\mathcal{W}_h = \text{span}(\{1, x, x^2, x^3\})$$

so that if $w_h \in \mathcal{W}_h$, then $w_h = w_0 \cdot 1 + w_1 x + w_2 x^2 + w_3 x^3$. For this example, we are going to seek a function u_h that satisfy a Dirichlet boundary condition at $x = 0$, in this case $u_h(0) = 2$, so we will set the trial space to

$$\begin{aligned}\mathcal{S}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = 2\} \\ &= \{w_h = 2 + w_1 x + w_2 x^2 + w_3 x^3 \mid (w_1, w_2, w_3) \in \mathbb{R}^3\}.\end{aligned} \tag{1.67}$$

For $\mathcal{V}_h$, we find the direction of $\mathcal{S}_h$ to get

$$\begin{aligned}\mathcal{V}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = 0\} \\ &= \text{span}(\{x, x^2, x^3\}).\end{aligned}$$

The method consists in finding $u_h \in \mathcal{S}_h$ such that

$$a(u_h, v_h) \int_0^1 u_h' v_h' \, dx = \int_0^1 v_h \, dx = \ell(v_h) \tag{1.68}$$

for all $v_h \in \mathcal{V}_h$.

We can now find $u_h = 2 + u_1 x + u_2 x^2 + u_3 x^3$. To this end, we will find u_1, u_2 and u_3 by testing the variational equation with each function in the basis for $\mathcal{V}_h$. To wit,

$$\begin{aligned}\int_0^1 (u_1 + 2u_2 x + 3u_3 x^2)\, dx &= \int_0^1 x \, dx \\ \int_0^1 (u_1 + 2u_2 x + 3u_3 x^2) 2x \, dx &= \int_0^1 x^2 \, dx \\ \int_0^1 (u_1 + 2u_2 x + 3u_3 x^2) 3x^2 \, dx &= \int_0^1 x^3 \, dx.\end{aligned}$$

Evaluating the integrals, we obtain

$$\begin{aligned} u_1 + u_2 + u_3 &= \frac{1}{2} \\ u_1 + \frac{4}{3}u_2 + \frac{3}{2}u_3 &= \frac{1}{3} \\ u_1 + \frac{3}{2}u_2 + \frac{9}{5}u_3 &= \frac{1}{4}. \end{aligned}$$

This defines a system of 3 equations with 3 unknonws, u_1, u_2 and u_3. The solution is $u_1 = 1$, $u_2 = -1/2$ and $u_3 = 0$. Therefore,

$$u_h = 2 + x - x^2/2.$$

1.3.1.1 Choice of Trial and Test Spaces

In §1.1.2 we showed that the solution u of a problem such as Problem 1.1 satisfies different variational equations. Each variational equation separates the boundary conditions in Problem 1.1 into two classes, essential and natural boundary conditions. The same boundary condition can be essential for one variational equation, and natural for another. For example, the Dirichlet boundary condition (1.6a) is an essential boundary condition for variational equation (1.9a) that we obtain from the recipe in §1.1.2.3, but it is a natural boundary condition for the variational equation in Nitsche's method, (1.31). Therefore, boundary conditions are not intrinsically essential or natural, but they acquire that role for a given variational equation.

When approximating the solution u of a problem such as Problem 1.1 with a variational numerical method, we will first need to identify whether the boundary conditions of the problem are essential or natural for the variational equation defined by F (or by a and ℓ in Problem 1.2). Then, any essential boundary condition will need to be enforced through the choice of the trial space $\mathcal{S}_h$. The simplest way to enforce them is to require that any function that belongs to $\mathcal{S}_h$ satisfy the essential boundary conditions. Natural boundary conditions are going to be enforced by the variational equation.

Later in §YYY, we will see that in many situations we will need to require functions in $\mathcal{S}_h$ to satisfy essential boundary conditions only approximately, since it is not possible to exactly enforce them.

Given a variational equation $F(u, v) = 0$ for all $v \in \mathcal{V}$ that u satisfies, a variational method for it is:

$$\text{Find } u_h \in \mathcal{S}_h \text{ such that } F(u_h, v_h) = 0 \text{ for all } v_h \in \mathcal{V}_h.$$

In this case, it is convenient to also require that $\mathcal{V}_h \subseteq \mathcal{V}$, so that it also holds that

$$F(u, v_h) = 0 \qquad \forall v_h \in \mathcal{V}_h. \tag{1.69}$$

In other words, the exact solution satisfies the variational equations used in the numerical method. A method that satisfies (1.69) is said to be **consistent**, and (1.69) is called a **consistency condition**. This condition will play a crucial role to

guarantee that we can approximate u with the method, as we will the have the chance to discuss in §3.

Summarizing,

- $\mathcal{S}_h = \{u_h \in \mathcal{W}_h \mid u_h \text{ satisfies essential boundary conditions for } F\}$
- $\mathcal{V}_h$ = Direction of $\mathcal{S}_h$
- For consistency, we require $\mathcal{V}_h \subseteq \mathcal{V}$.

We will illustrate these ideas by considering three different methods for Problem 1.1 with $k(x) = 1$, $b(x) = c(x) = 0$. That is, the problem is to find $u\colon [0, L] \to \mathbb{R}$ that satisfies that

$$-u''(x) = f(x) \qquad x \in (0, L) \tag{1.70a}$$
$$u(0) = g_0 \tag{1.70b}$$
$$u'(L) = d_L. \tag{1.70c}$$

For concreteness, we set

$$\mathcal{W}_h = \operatorname{span}(1, x, x^2, x^3),$$

so that we can provide explicit expressions for $\mathcal{S}_h$ and $\mathcal{V}_h$. This is the same space we adopted in Example 1.51. In the three example, the method reads:

$$\text{Find } u_h \in \mathcal{S}_h \text{ such that } a_h(u_h, v_h) = \ell(v_h) \text{ for all } v_h \in \mathcal{V}_h.$$

We indicate choices for a_h, ℓ_h, $\mathcal{V}_h$ and $\mathcal{S}_h$.

Examples:

1.52 The most common variational method adopts variational equation (1.16), so

$$a(u_h, v_h) = \int_0^L u_h' v_h' \, dx$$
$$\ell(v_h) = \int_0^L f v_h \, dx + d_L v_h(L).$$

For this variational equation, the Dirichlet boundary condition is essential, so we set

$$\begin{aligned}\mathcal{S}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = g_0\} \\ &= \{g_0 + c_1 x + c_2 x^2 + c_3 x^3 \mid (c_1, c_2, c_3) \in \mathbb{R}^3\}\end{aligned}$$

and its direction is

$$\begin{aligned}\mathcal{V}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = 0\} \\ &= \{c_1 x + c_2 x^2 + c_3 x^3 \mid (c_1, c_2, c_3) \in \mathbb{R}^3\}.\end{aligned}$$

Because functions in $\mathcal{V}_h$ are smooth, $\mathcal{V}_h \subset \mathcal{V} = \{w\colon [0,L] \to \mathbb{R} \text{ smooth } | \ w(0) = 0\}$, and the method is consistent. This is the method we used in Example 1.51.

1.53 **Nitsche's Method.** For this also the method we adopt variational equation (1.31), so for $\mu > 0$,

$$a(u_h, v_h) = \int_0^L u_h' v_h' \, dx + u_h'(0)\, v_h(0) - u_h(0)\, v_h'(0) + \mu u_h(0)\, v_h(0)$$
$$\ell(v_h) = \int_0^L f v_h \, dx + d_L v_h(L) - g_0 v_h'(0) + \mu g_0 v_h(0).$$

All boundary conditions are natural, so we can set

$$\mathcal{S}_h = \mathcal{V}_h = \mathcal{W}_h = \{c_0 + c_1 x + c_2 x^2 + c_3 x^3 \mid (c_0, c_1, c_2, c_3) \in \mathbb{R}^4\}.$$

Since $\mathcal{W}_h$ is a vector space, it is also its own direction. Since $\mathcal{V}_h \subset \mathcal{V} = \{w\colon [0,L] \to \mathbb{R} \text{ smooth}\}$, the method is consistent.

1.54 In this case, we adopt variational equation (1.18), so we set

$$a(u_h, v_h) = -\int_0^L u'' v_h \, dx$$
$$\ell(v_h) = \int_0^L f v_h \, dx.$$

As mentioned in Example 1.12, for this variational equation 1.18 both boundary conditions are essential. Therefore, we need to require them in the definition of $\mathcal{S}_h$:

$$\begin{aligned}\mathcal{S}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = g_0, w_h'(L) = d_L\} \\ &= \{g_0 + c_1 x + c_2 x^2 + c_3 x^3 \mid c_1 + 2c_2 L + 3c_3 L^2 = d_L, (c_1, c_2, c_3) \in \mathbb{R}^3\}.\end{aligned}$$

Its direction is

$$\begin{aligned}\mathcal{V}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = 0, w_h'(L) = 0\} \\ &= \{c_1 x + c_2 x^2 + c_3 x^3 \mid c_1 + 2c_2 L + 3c_3 L^2 = 0, (c_1, c_2, c_3) \in \mathbb{R}^3\}.\end{aligned}$$

These two spaces have a non-trivial constraint in their definition, but it is simple to solve for c_1 in both cases and replace in the expression for the functions.

The test space of (1.18) is

$$\mathcal{V} = \{w\colon [0,L] \to \mathbb{R}\},$$

so $\mathcal{V}_h \subset \mathcal{V}$, and the method is consistent.

1.3.1.2 Continuous/Discontinuous Galerkin and Petrov-Galerkin Methods

Variational methods are commonly referred to with names that originate on features of the trial and test spaces:

Bubnov-Galerkin, Galerkin, or Continuous Galerkin Method: Functions in the affine space $\mathcal{S}_h$ are continuous, and $\mathcal{V}_h$ is the direction of $\mathcal{S}_h$.

Discontinuous Galerkin Method: Functions in the affine space $\mathcal{S}_h$ can be discontinuous, and $\mathcal{V}_h$ is the direction of $\mathcal{S}_h$.

Petrov-Galerkin Method: The test space $\mathcal{V}_h$ is *not* the direction of trial space $\mathcal{S}_h$.

This classification of finite element methods should be considered a guideline and not a definition, since the literature has blurry boundaries on what constitutes, for example, a Galerkin method.

In his original work, Boris Grigoryevich Galerkin [3] proposed a method in which the spaces $\mathcal{S}_h$ and $\mathcal{V}_h$ were one and the same, Dirichlet and Neumann boundary conditions were satisfied by all functions in $\mathcal{V}_h$, and the variational equation was obtained without integrating by parts, or the method of weighted residuals in (1.22). In his landmark paper, Galerkin cites W. Ritz [9, 8], whose method relied on the existence of a potential energy to be minimized, instead of solely a differential equation to be enforced. This is why Galerkin method is also often referred to as **Ritz-Galerkin Method**. Over time, the name Galerkin method has been adopted for to denominate extensions and generalizations of the original method.

1.3.2 Solution to a Variational Method

We next describe the general procedure to find the solution of (1.66), regardless of the way we construct the discrete space $\mathcal{W}_h$.

Let $\{N_a\}_{a=1,\dots,m} = \{N_1, \dots N_m\}$, $m \in \mathbb{N}$, be a basis for $\mathcal{W}_h$. Then, the approximate solution $u_h \in \mathcal{S}_h$ of Problem 1.2 and any test function $v_h \in \mathcal{V}_h$ can be written as

$$u_h(x) = \sum_{b=1}^{m} u_b N_b(x)$$
$$v_h(x) = \sum_{a=1}^{m} v_a N_a(x).$$

Additionally, we will assume that the subset of basis functions $\{N_a\}_{a=1,\dots,n}$ with $n \leq m$ is a basis for $\mathcal{V}_h$. Graphically,

$$\underbrace{\overbrace{N_1, \dots, N_n}^{\text{Basis for } \mathcal{V}_h}, \dots, N_m}_{\text{Basis for } \mathcal{W}_h}. \tag{1.71}$$

This automatically means that $v_a = 0$ for $n < a \leq m$.

The solution u_h satisfies (1.66) for any $v_h \in \mathcal{V}_h$. To find u_h, we will take advantage that we can choose the test functions v_h we can "test" with, and of the fact that u_h belongs to $\mathcal{S}_h$. If we choose enough test functions, we will get enough equations to define u_h completely. We can then show that such u_h satisfies (1.66) for any $v_h \in \mathcal{V}_h$, not only for those chosen as particular test functions.

For this plan, we will select each basis function of $\mathcal{V}_h$ as a test function, namely,

$$\ell(N_a) = a(u_h, N_a) \qquad a = 1, \dots, n. \tag{1.72a}$$

This gives us n algebraic equations, for the m unknown components $\{u_1, \dots, u_m\}$ of u in the basis $\{N_a\}_{a=1,\dots,m}$. The remaining $n-m$ equations follow from the fact that $u_h \in \mathcal{S}_h$. Typically, this means that the remaining equations come from the boundary conditions. To impose them, it is enough to select *any* element $\overline{u}_h$ of $\mathcal{S}_h$, write

$$\overline{u}_h = \underbrace{\overline{u}_1 N_1 + \dots + \overline{u}_n N_n}_{\in \mathcal{V}_h} + \underbrace{\dots + \overline{u}_m N_m}_{\notin \mathcal{V}_h}$$

and set

$$u_a = \overline{u}_a \qquad n < a \le m, \tag{1.72b}$$

which provide the remaining $n-m$ equations needed to completely determine the m components $u_1, \dots, u_m$ of u_h in the basis $\{N_1, \dots, N_m\}$.

The solution to (1.72) amounts to the solution of a linear system of equations. To see this, we first expand u_h in components inside (1.72a) and use the bilinearity of a to get:

$$\ell(N_a) = a(u_h, N_a) = a\left(\sum_{b=1}^{m} u_b N_b, N_a\right) = \sum_{b=1}^{m} a(N_b, N_a)\, u_b \qquad a = 1, \dots, n. \tag{1.73}$$

We then label

$$F_a = \ell(N_a), \qquad K_{ab} = a(N_b, N_a), \qquad 1 \le a \le n,\ 1 \le b \le m$$

and from (1.72b),

$$F_a = \overline{u}_a, \qquad K_{ab} = \delta_{ab}, \qquad n < a \le m,\ 1 \le b \le m$$

where δ_{ab} is called the **Kronecker Delta** [3], and arrange them in a matrix and two columns vectors

$$K = \begin{bmatrix} K_{11} & K_{12} & \dots & K_{1m} \\ K_{21} & K_{22} & \dots & K_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ K_{m1} & K_{m2} & \dots & K_{mm} \end{bmatrix}, \quad F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}, \text{ and } U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}. \tag{1.74}$$

[3] It is defined as $\delta_{ab} = \begin{cases} 1 & a = b, \\ 0 & a \neq b. \end{cases}$

☞ We could be a little bit more specific about K, and use that we know the values for $a, b > n$. Namely,

$$K = \begin{bmatrix} K_{11} & \dots & K_{1(n+1)} & \dots & K_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ K_{n1} & \dots & K_{n(n+1)} & \dots & K_{nm} \\ 0 & \dots & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 1 \end{bmatrix}$$

Then, (1.72) is expressed as the linear system of equations

$$KU = F. \tag{1.75}$$

The matrix K is often called the **stiffness matrix** and the vector F is often called the **load vector**, for their origins in mechanical problems.

Solving the linear system (1.75) for U defines the components $u_1, \ldots, u_m$ needed to construct the function $u_h = u_1 N_1 + \ldots + u_m N_m$, the solution to the variational method.

Example 1.55 Let's revisit example 1.51, to see that we have solved it exactly as we outlined here. The spaces in the example are:

$$\begin{aligned} \mathcal{W}_h &= \text{span}(\{1, x, x^2, x^3\}) \\ \mathcal{V}_h &= \text{span}(\{x, x^2, x^3\}) \\ \mathcal{S}_h &= \{2 + v_h \mid v_h \in \mathcal{V}_h\}. \end{aligned}$$

Hence, we have $m = 4$ and $n = 3$, and index the basis functions in $\mathcal{W}_h$ so that indices 1 to 3 form a basis for $\mathcal{V}_h$. To wit, we set $N_1(x) = x$, $N_2(x) = x^2$, $N_3(x) = x^3$ and $N_4(x) = 1$. We also need to choose one element $\overline{u}_h$ of $\mathcal{S}_h$. Among the infinite options we have here, one is $\overline{u}_h(x) = 2N_4(x) = 2$, and another one is $\overline{u}_h(x) = 3N_1(x) + 2N_4(x)$. Notice that regardless of what we choose for $\overline{u}_h$, all of them will have $\overline{u}_4 = 2$, since this is the only way to construct the constant function 2 needed to belong to $\mathcal{S}_h$.

With these choices, the equations imposed by the variational method, (1.72a), are

$$\begin{aligned} a(u_h, N_1) &= \ell(N_1) \\ a(u_h, N_2) &= \ell(N_2) \\ a(u_h, N_3) &= \ell(N_3) \end{aligned}$$

while the equations that impose that $u_h \in \mathcal{S}_h$, (1.72b), is

$$u_4 = 2.$$

Replacing, the load vector is

$$F = \begin{bmatrix} \ell(N_1) \\ \ell(N_2) \\ \ell(N_3) \\ u_4 \end{bmatrix} = \begin{bmatrix} \int_0^1 x\,dx \\ \int_0^1 x^2\,dx \\ \int_0^1 x^3\,dx \\ \overline{u}_4 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/3 \\ 1/4 \\ 2 \end{bmatrix}.$$

The stiffness matrix is

$$
\begin{aligned}
K &= \begin{bmatrix} a(N_1,N_1) & a(N_2,N_1) & a(N_3,N_1) & a(N_4,N_1) \\ a(N_1,N_2) & a(N_2,N_2) & a(N_3,N_2) & a(N_4,N_2) \\ a(N_1,N_3) & a(N_2,N_3) & a(N_3,N_3) & a(N_4,N_3) \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} \int_0^1 1\cdot 1\,dx & \int_0^1 2x\cdot 1\,dx & \int_0^1 3x^2\cdot 1\,dx & \int_0^1 0\cdot 1\,dx \\ \int_0^1 1\cdot 2x\,dx & \int_0^1 2x\cdot 2x\,dx & \int_0^1 3x^2\cdot 2x\,dx & \int_0^1 0\cdot 2x\,dx \\ \int_0^1 1\cdot 3x^2\,dx & \int_0^1 2x\cdot 3x^2\,dx & \int_0^1 3x^2\cdot 3x^2\,dx & \int_0^1 0\cdot 3x^2\,dx \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 4/3 & 3/2 & 0 \\ 1 & 3/2 & 9/5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.
\end{aligned}
$$

The components of u_h are then

$$
U = K^{-1}F = \begin{bmatrix} 1 \\ -1/2 \\ 0 \\ 2 \end{bmatrix}, \tag{1.76}
$$

from where

$$
u_h(x) = 1.N_1(x) - 1/2N_2(x) + 0.N_3(x) + 2N_4(x) = 2 + x - \frac{x^2}{2}. \tag{1.77}
$$

This happens to be the *exact* solution of the problem, whose strong form consists of the following three equations:

$$
\begin{aligned}
-u''(x) &= 1 \qquad x \in (0,1) \\
u(0) &= 2 \\
u'(1) &= 0.
\end{aligned}
$$

This can be understood because in this case $\mathscr{S}_h$ *contains* the exact solution of the problem. A *consistent* variational method will *always* find the exact solution if it belongs to $\mathscr{S}_h$ and it is the sole solution of Problem 1.2. In general, however, this will not be the case.

1.3.2.1 Why does this solution procedure work?

We complete the last section by answering two questions about this solution procedure. The first question is:

Why do we test with the basis functions only, if variational equation (1.66) *should hold for* all *test functions?*

The answer is that if the variational equation is satisfied for every function in a basis for the test space $\mathcal{V}_h$, it is satisfied for *every* function in the test space.

The proof is simple, and worth reading, and it takes advantage of the bilinearity of a and the linearity of ℓ:

$$\begin{aligned} a(u_h, v_h) &= a\left(u_h, \sum_{b=1}^{n} v_b N_b\right) \\ &= \sum_{b=1}^{n} v_b a(u_h, N_b) && \text{bilinearity of } a \\ &= \sum_{b=1}^{n} v_b \ell(N_b) && \text{use of (1.72a)} \\ &= \ell\left(\sum_{b=1}^{n} v_b N_b\right) && \text{linearity of } \ell \\ &= \ell(v_h). \end{aligned}$$

So, for $u_h \in \mathcal{S}_h$, (1.72a) implies (1.66). The converse is trivially true, namely, if (1.66) is satisfied for any $v_h \in \mathcal{V}_h$, it is satisfied for any basis function $N_b \in \mathcal{V}_h$ in particular, and hence it implies (1.72a). In summary, if $u_h \in \mathcal{S}_h$,

$$u_h \text{ is a solution of (1.66)} \iff u_h \text{ is a solution of (1.72a)}.$$

In words, this implies that the solution of the variational method is a solution of the linear system of equations defined by the basis functions of the test space, and conversely[4].

The second question we answer is:

Why does the solution u_h belong to $\mathcal{S}_h$, and why is it independent of our choice of $\overline{u}_h \in \mathcal{S}_h$?

The answer to this relies on the fact that $\mathcal{S}_h$ is an affine subspace of $\mathcal{W}_h$ and $\mathcal{V}_h$ is its direction. To see that $u_h \in \mathcal{S}_h$, notice that since $u_b = \overline{u}_b$ for $n < b \le m$, then

$$\Delta u_h = u_h - \overline{u}_h = (u_1 - \overline{u}_1)N_1 + \ldots + (u_n - \overline{u}_n)N_n,$$

from where we conclude that $\Delta u_h \in \mathcal{V}_h$, or $u_h = \overline{u}_h + \Delta u_h$, and hence it follows from (1.47) that $u_h \in \mathcal{S}_h$. To see that the choice of $\overline{u}_h$ does not affect the u_h we compute, consider another function $\overline{w}_h \in \mathcal{S}_h$. Then, by the definition of affine subspace, $\overline{u}_h - \overline{w}_h \in \mathcal{V}_h$, or in terms of the basis for $\mathcal{W}_h$,

$$\overline{u}_h = \overline{w}_h + \sum_{b=1}^{n} v_b N_b.$$

So, $\overline{u}_h$ and $\overline{w}_h$ can only differ in the values of the components $v_1, \ldots, v_n$, but need to have precisely the same values for the components v_b with $n < b \le m$. Since the latter are the only components that participate in (1.72b), the solution u_h does not change if we choose $\overline{w}_h$ instead. In other words, u_h does not depend on our choice of $\overline{u}_h$.

[4] It is possible to regard (1.72a) as the Euler-Lagrange equations of (1.66).

1.3.2.2 Solution to a Variational Method With an Arbitrarily-Ordered Basis

In general, an ordered basis as in (1.71) is not readily available, nor is it necessary. We discuss next how to proceed in the case in which the basis functions for $\mathcal{W}_h$ and $\mathcal{V}_h$ are not neatly ordered as in the earlier discussion.

Again, let $\{N_a\}_{a=1,\dots,m}$ be a basis for $\mathcal{W}_h$, and again we will assume that a subset of $n \le m$ of these basis functions is a basis for $\mathcal{V}_h$. However, the basis for $\mathcal{V}_h$ need *not* be the set $\{N_a\}_{a=1,\dots,n}$. To indicate the basis for $\mathcal{V}_h$, it is convenient to introduce three sets of indices, or **index sets**. First, we denote by $\eta = \{1,\dots,m\}$ the set of indices of all basis functions in $\mathcal{W}_h$. The basis functions for $\mathcal{V}_h$ can be indicated by a subset of η. The set of indices of basis functions for $\mathcal{V}_h$ is denoted $\eta_a \subset \eta$; it is called the set of **active indices**, and we can write

$$\mathcal{V}_h = \text{span}\left(\bigcup_{a\in\eta_a} \{N_a\} \right)$$

or

$$w_h \in \mathcal{V}_h \iff w_h = \sum_{a\in\eta_a} w_a N_a.$$

The remaining indices in η, those that are *not* in η_a, is denoted $\eta_g = \eta \setminus \eta_a$; it is called the set of **constrained indices**.

We next rewrite the equations to solve the variational method using these index sets. First, testing with each basis function in $\mathcal{V}_h$, (1.72a), is restated as

$$\ell(N_a) = a(u_h, N_a) \qquad a \in \eta_a \tag{1.78a}$$

The arbitrary element $\overline{u}_h \in \mathcal{S}_h$ used to impose the fact that $u_h \in \mathcal{S}_h$ is still written as $\overline{u}_h = \overline{u}_1 N_1 + \dots + \overline{u}_m N_m$, but (1.72b) is restated as

$$u_a = \overline{u}_a \qquad a \in \eta_g. \tag{1.78b}$$

We then label

$$\begin{aligned} F_a &= \ell(N_a), & K_{ab} &= a(N_b, N_a) & a &\in \eta_a, b \in \eta \\ F_a &= \overline{u}_a, & K_{ab} &= \delta_{ab} & a &\in \eta_g, b \in \eta \end{aligned} \tag{1.78c}$$

which define the stiffness matrix K and load vector F.

To illustrate these ideas, let's consider Example 1.55 again.

Example 1.56 Consider Example 1.55 again, but in this case we set $N_1(x) = x$, $N_2(x) = 1$, $N_3(x) = x^2$ and $N_4(x) = x^3$. Therefore, the basis for $\mathcal{V}_h$ is $\{N_1, N_3, N_4\}$, and the index sets are $\eta = \{1,2,3,4\}$, $\eta_a = \{1,3,4\}$, and $\eta_g = \{2\}$. We can then set $\overline{u}_h = 2N_2(x) = 2$.

The stiffness matrix and load vector in this case are

$$F = \begin{bmatrix} \int_0^1 x\,dx \\ \overline{u}_2 \\ \int_0^1 x^2\,dx \\ \int_0^1 x^3\,dx \end{bmatrix} = \begin{bmatrix} 1/2 \\ 2 \\ 1/3 \\ 1/4 \end{bmatrix}.$$

The stiffness matrix is

$$K = \begin{bmatrix} \int_0^1 1\cdot 1\,dx & \int_0^1 0\cdot 1\,dx & \int_0^1 2x\cdot 1\,dx & \int_0^1 3x^2\cdot 1\,dx \\ 0 & 1 & 0 & 0 \\ \int_0^1 1\cdot 2x\,dx & \int_0^1 0\cdot 2x\,dx & \int_0^1 2x\cdot 2x\,dx & \int_0^1 3x^2\cdot 2x\,dx \\ \int_0^1 1\cdot 3x^2\,dx & \int_0^1 0\cdot 3x^2\,dx & \int_0^1 2x\cdot 3x^2\,dx & \int_0^1 3x^2\cdot 3x^2\,dx \end{bmatrix}$$
$$= \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 4/3 & 3/2 \\ 1 & 0 & 3/2 & 9/5 \end{bmatrix}.$$

The components of the solution are

$$U = \begin{bmatrix} 1 \\ 2 \\ -1/2 \\ 0 \end{bmatrix},$$

and the solution is

$$u_h(x) = 1.N_1(x) + 2N_2(x) - 1/2N_3(x) + 0N_4(x) = 2 + x - \frac{x^2}{2}, \tag{1.79}$$

which is exactly the same function we obtained in Example 1.55.

Comparing the stiffness matrix and load vector in Examples 1.55 and 1.56, we notice that they have the same entries, but reordered: the last row and column in Example 1.55 were moved to be the second row and column in Example 1.56. The solution U in the former has the last row moved to the be the second row in the latter. Of course, the solution u_h is the same in both cases, since the entries in U are multiplied by the reordered basis functions as well.

To conclude this discussion, notice that reordering the basis functions does not change the spaces $\mathcal{V}_h$, $\mathcal{S}_h$ and $\mathcal{W}_h$, and hence it should not change the solution to the variational method.

Example 1.57 Let's look at another example of a variational method, in this case with a basis of trigonometric functions. To this end, we will revisit Example 1.8 in a domain $\Omega = [0, \pi/2]$. The problem is given by

$$-u''(x) + u'(x) + u(x) = -5\exp(-2x), \qquad \forall\, x \in \Omega, \tag{1.80a}$$
$$u(0) = 1, \tag{1.80b}$$
$$u'(\pi/2) = -2\exp(-\pi). \tag{1.80c}$$

and the exact solution is $u(x) = \exp(-2x)$. A variational equation that the solution u satisfies is

$$\int_0^{\pi/2} u'(x)v'(x) + u'(x)v(x) + u(x)v(x)\,dx + 2\exp(-\pi)v(\pi/2) = -\int_0^{\pi/2} 5\exp(-2x)v(x)\,dx, \tag{1.80d}$$

where

$$\mathcal{V} = \{w\colon [0,1] \to \mathbb{R} \text{ smooth } \mid w(0) = 0\}.$$

The bilinear form and linear functional here are

$$a(u,v) = \int_0^{\pi/2} u'(x)v'(x) + u'(x)v(x) + u(x)v(x)\,dx \tag{1.80e}$$

$$\ell(v) = -\int_0^{\pi/2} 5\exp(-2x)v(x)\,dx - 2\exp(-\pi)v(\pi/2). \tag{1.80f}$$

The bilinear form is not symmetric. In this variational equation, boundary condition (1.80b) is essential and (1.80c) is natural.

To formulate a variational method, we consider the space

$$\mathcal{W}_h^n = \text{span}(\{1, \sin x, \ldots, \sin nx\})$$

for $n \in \mathbb{N}$. We included a dependence on n for generality, but we will proceed with $n = 2$ next. To this end, we will label $N_1(x) = 1, N_2(x) = \sin x, N_3(x) = \sin 2x$.

Let's find spaces $\mathcal{S}_h$ and $\mathcal{V}_h$. For $w_h \in \mathcal{W}_h^2$, we can write

$$\begin{aligned} w_h(x) &= w_1 N_1(x) + w_2 N_2(x) + w_3 N_3(x) \\ &= w_1.1 + w_2 \sin x + w_3 \sin 2x. \end{aligned}$$

The space $\mathcal{S}_h$ follows by requiring essential boundary condition (1.80b) to be satisfied by the functions in it, namely,

$$\begin{aligned} \mathcal{S}_h &= \{w_h \in W_h^2 \mid w_h(0) = 1\} \\ &= \{1 + w_2 \sin x + w_3 \sin 2x \mid (w_2, w_3) \in \mathbb{R}^2\} \\ \mathcal{V}_h &= \{w_h \in W_h^2 \mid w_h(0) = 0\} \\ &= \{w_2 \sin x + w_3 \sin 2x \mid (w_2, w_3) \in \mathbb{R}^2\}. \end{aligned}$$

Here $\mathcal{V}_h$ is the direction of $\mathcal{S}_h$, and $\mathcal{V}_h \subset \mathcal{V}$, so the method is consistent.

To proceed, we need to identify active and constrained indices. In this case, $\eta_c = \{1\}$ and $\eta_a = \{2,3\}$. The stiffness matrix is then (careful because this is a non-symmetric bilinear form):

$$K = \begin{bmatrix} 1 & 0 & 0 \\ a(N_1,N_2) & a(N_2,N_2) & a(N_3,N_2) \\ a(N_1,N_3) & a(N_2,N_3) & a(N_3,N_3) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & (1+\pi)/2 & 2/3 \\ 1 & 2 & 5\pi/4 \end{bmatrix}.$$

We set $\overline{u}_h(x) = 1$, so that $\overline{u}_h \in \mathcal{S}_h$, and $u_1 = 1$. The load vector is

$$F = \begin{bmatrix} 1 \\ \ell(N_2) \\ \ell(N_3) \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ -5(1+\exp(-\pi))/4 \end{bmatrix}.$$

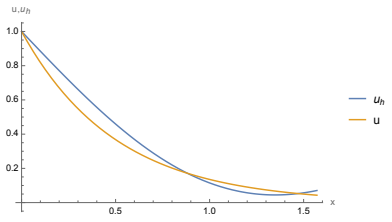


Figure 1.6

The components of the solution are obtained from $U = K^{-1}F$, or

$$U = \begin{bmatrix} 1 \\ -\frac{36+20\exp(-\pi)-60\pi}{32-15\pi-15\pi^2} \\ -\frac{3\exp(-\pi)\left(5(1+\pi)+\exp(\pi)(9\pi-23)\right)}{-32+15\pi+15\pi^2} \end{bmatrix} \approx \begin{bmatrix} 1 \\ -0.93 \\ -0.11 \end{bmatrix}.$$

Hence,

$$u_h(x) \approx 1 - 0.93\sin x - 0.11\sin 2x.$$

A plot of the exact versus the approximate solutions is shown in Fig. 1.6. By selecting a larger value of n, a better approximation is obtained. You can check that.

Example 1.58 Let's consider a twist of Example 1.57 to illustrate the effect of additional boundary conditions on the method. To this end, we change the problem in that example to have a Dirichlet boundary condition at $x = \pi/2$ as well, keeping the same exact solution. The problem is

$$-u''(x) + u'(x) + u(x) = -5\exp(-2x), \qquad \forall\, x \in \Omega, \tag{1.81a}$$

$$u(0) = 1, \tag{1.81b}$$

$$u(\pi/2) = \exp(-\pi). \tag{1.81c}$$

and the exact solution is still $u(x) = \exp(-2x)$. The variational equation that the solution u of this problem satisfies is:

$$\int_0^{\pi/2} u'(x)\,v'(x) + u'(x)\,v(x) + u(x)\,v(x)\,dx = -\int_0^{\pi/2} 5\exp(-2x)\,v(x)\,dx \qquad \forall\, v \in \mathcal{V}, \tag{1.81d}$$

where

$$\mathcal{V} = \{w\colon [0,1] \to \mathbb{R} \text{ smooth } \mid w(0) = 0, w(\pi/2) = 0\}.$$

Notice that there is no longer a term that appears from the natural boundary condition. Both boundary conditions are now essential.

The bilinear form and linear functional here are

$$a(u,v) = \int_0^{\pi/2} u'(x)\,v'(x) + u'(x)\,v(x) + u(x)\,v(x)\,dx \tag{1.81e}$$

$$\ell(v) = -\int_0^{\pi/2} 5\exp(-2x)\,v(x)\,dx. \tag{1.81f}$$

The bilinear form is not symmetric.

In this case, we consider the space

$$\mathcal{W}_h = \text{span}(\{1, \sin x, \sin 2x, \sin 4x\}).$$

We will label $N_1(x) = 1, N_2(x) = \sin x, N_3(x) = \sin 2x,\ N_4(x) = \sin 4x$. Notice that we did not include the function $\sin 3x$.

Let's find spaces $\mathcal{S}_h$ and $\mathcal{V}_h$ next. For $w_h \in \mathcal{W}_h$, we can write

$$\begin{aligned} w_h(x) &= w_1 N_1(x) + w_2 N_2(x) + w_3 N_3(x) + w_4 N_4(x) \\ &= w_1.1 + w_2 \sin x + w_3 \sin 2x + w_4 \sin 4x. \end{aligned}$$

For $w_h \in \mathcal{S}_h$, we need $1 = w_h(0) = w_1$ and $\exp(-\pi) = w_h(\pi/2) = w_1 + w_2$, or $w_1 = 1$ and $w_2 = \exp(-\pi) - 1$. Therefore,

$$\begin{aligned} \mathcal{S}_h &= \{w_h \in W_h \mid w_h(0) = 1, w_h(\pi/2) = \exp(-\pi)\} \\ &= \{1 + (\exp(-\pi) - 1)\sin x + w_3 \sin 2x + w_4 \sin 4x \mid (w_3, w_4) \in \mathbb{R}^2\} \\ \mathcal{V}_h &= \{w_h \in W_h \mid w_h(0) = w_h(\pi/2) = 0\} \\ &= \{w_3 \sin 2x + w_4 \sin 4x \mid (w_3, w_4) \in \mathbb{R}^2\}. \end{aligned}$$

Had we included a term with $\sin 3x$, the characterization of $\mathcal{S}_h$ and $\mathcal{V}_h$ would have been somewhat more complicated, because we would have had a total of 3 functions that are non-zero at $x = \pi/2$.

The active and constrained indices are $\eta_c = \{1, 2\}$ and $\eta_a = \{3, 4\}$. The stiffness matrix is then (careful because this is a non-symmetric bilinear form):

$$\begin{aligned} K &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ a(N_1, N_3) & a(N_2, N_3) & a(N_3, N_3) & a(N_4, N_3) \\ a(N_1, N_4) & a(N_2, N_4) & a(N_3, N_4) & a(N_4, N_4) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 2 & 5\pi/4 & -4/3 \\ 0 & -4/15 & 4/3 & 17\pi/4 \end{bmatrix}. \end{aligned}$$

We set $\overline{u}_h(x) = 1 + (\exp(-\pi) - 1)\sin x$, so that $\overline{u}_h \in \mathcal{S}_h$, $u_1 = 1$ and $u_2 = \exp(-\pi) - 1$. The load vector is

$$F = \begin{bmatrix} 1 \\ \exp(-\pi) - 1 \\ \ell(N_3) \\ \ell(N_4) \end{bmatrix} = \begin{bmatrix} 1 \\ \exp(-\pi) - 1 \\ 5(1 + \exp(-\pi))/4 \\ \exp(-\pi) - 1 \end{bmatrix}.$$

The components of the solution are obtained from $U = K^{-1}F$, or

$$U = \begin{bmatrix} 1 \\ \exp(-\pi) - 1 \\ \frac{\exp(-\pi)(-e^{\pi}(1216 + 765\pi) - 9945\pi + 1216)}{5(256 + 765\pi^2)} \\ \frac{12\exp(-\pi)(e^{\pi}(4 - 19\pi) + 19\pi + 52)}{256 + 765\pi^2} \end{bmatrix} \approx \begin{bmatrix} 1 \\ -0.96 \\ -0.13 \\ -0.08 \end{bmatrix}.$$

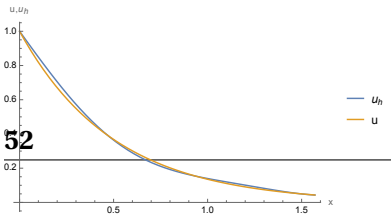


Figure 1.7

The solution is then

$$1 - 0.96 \sin x - 0.13 \sin 2x - 0.08 \sin 4x,$$

and it is plotted in Fig. 1.7.

1.3.3 The Euler-Lagrange Equations

Imagine that a friend described a variational equation to you that a function u should satisfy. Could you find out what differential equation it corresponds to, if any, and what boundary conditions the variational equation requires u to satisfy, if any? In other words, does the fact that u satisfies a variational equation imply that it should also satisfy a differential equation and/or some boundary conditions? By definition, such boundary conditions are what we have termed natural boundary conditions for u.

For example, in engineering many problems require finding the minimizer of a functional, the most well-known one is minimizing the potential energy of a physical system. Minimizing a functional often leads to a variational equation that the minimizer needs to satisfy, and from here to a differential equation and/or boundary conditions.

You have already been in this situation: In examples 1.14, 1.15, and 1.16 we stated variational equations (1.31), (1.34) and (1.38) which we did not derive. How do we check if they imply what is stated in the example? For example, that a boundary condition is natural or essential. The process by which we answer this question will also be useful in §3 when we perform analysis of the convergence of the finite element method.

The first question to ask is how could the variational equation define a differential equation or boundary conditions for u. An intuitive argument could be made based on the finite dimensional case, or the variational method. In that case, by testing with all the basis functions in $\mathcal{V}_h$, we could determine u_h from the variational equation. Imagine now that we test with all functions in $\mathcal{V}$, can we determine a function u, up to perhaps the essential boundary conditions? We will see that often this will be the case; the variational equation can define a differential equation that the function u should satisfy.

When we talk about *conditions* imposed by a variational equation, we mean the following: For $u\colon \Omega \to \mathbb{R}$ that satisfies variational equation $F(u,v) = 0$ for all $v \in \mathcal{V}$, find a functional EL such that

$$F(u,v) = 0 \quad \forall v \in \mathcal{V} \qquad \Longleftrightarrow \qquad EL(u,x) = 0 \quad \forall x \in \omega \subseteq \overline{\Omega}, \tag{1.82}$$

Here ω is a subset of the closure of Ω, and as such, it may include points in both the interior and the boundary of Ω.

In the context of the Calculus of Variations, $EL(u,x) = 0$ for $x \in \omega$ is the Euler-Lagrange equation of the variational principle.

So, not only do we want to find conditions implied by the variational equation (left-to-right implication), but we want enough conditions that together imply

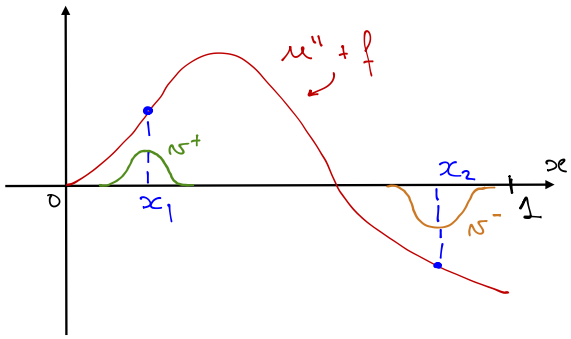


Figure 1.8 Potential choice of weighting functions to show that a weak solution is a strong solution.

that u satisfies the variational equation (right-to-left implication). Equation $EL(u,x) = 0 \quad \forall x \in \omega$ is called the **Euler-Lagrange** equation of the variational equation.

Let's illustrate how we do this through the simplest example. Assume that u satisfies variational equation (1.16). First, we integrate by parts the left hand side of (1.16) to remove the derivative of v', to get

$$0 = (u'(L) - d_L)\, v(L) - u'(0)\, v(0) - \int_0^L (u''(x) + f(x))\, v(x)\, dx. \tag{1.83}$$

Next, we use that we can choose any $v \in \mathcal{V}$. Referring to Fig. 1.8, we proceed by contradiction and assume that $u''(x_1) + f(x_1) > 0$ for some $x_1 \in (0, L)$. Then we can choose $v^+ \in \mathcal{V}$ as sketched in the figure[5]. In this case, using that $v^+(L) = v^+(0) = 0$, (1.83) reads

$$0 = -\int_0^L (u''(x) + f(x))\, v^+(x)\, dx. \tag{1.84}$$

But $v^+(x)$ is not negative anywhere, it is zero wherever $u''(x) + f(x)$ is negative, and both $u''(x) + f(x)$ and v^+ are positive in a neighborhood of x_1, so the integral in the right hand side of (1.84) needs to be positive, a contradiction. We conclude then that $u''(x) + f(x)$ cannot be positive anywhere. A similar argument can be made around a point in which $u'' + f$ is negative by selecting a weighting function v^-, as in the figure. We can then conclude that $u'' + f$ is neither positive nor negative anywhere in $(0, L)$, so $u''(x) + f(x) = 0$ for any $x \in (0, L)$. The function u then needs to satisfy (1.12a).

An important detail that often gets lost in a first view of the last argument is that each one of the different v's described above imposes a different condition

☞ It emerges from the discussion here that the set $\mathcal{V}$ should at least contain all functions v^+ and v^- for this argument to be made; the precise conditions for what functions $\mathcal{V}$ should include take the form of the "Fundamental lemma of the calculus of variations," see e.g. [4].

[5] For example, we can set

$$v^+(x) = \begin{cases} 0 & \text{if } |x - x_1| > \epsilon \\ \left[1 - \left(\frac{x - x_1}{\epsilon}\right)^2\right]^3 & \text{if } |x - x_1| \le \epsilon, \end{cases}$$

where $\epsilon > 0$ is the "half-width" of v^+, and can be chosen as small as desired.

☞ Often the following question is asked: "Wait... shouldn't u satisfy (1.12a) *only* when we consider those functions $v \in \mathcal{V}$ in Fig. 1.8? This question does not merit further consideration after realizing that whether u satisfies the differential equation or not depends on

on u, and that the only way for u to satisfy them all is by satisfying the differential equation (1.12a). Therefore, even though we may not have considered *all* functions $v \in \mathcal{V}$, we considered enough of them to conclude that u satisfies the differential equation.

The story is not over, as we see next. Since we just concluded that u needs to satisfy (1.12a), we may use it to simplify the variational equation, which reads

$$0 = (u'(L) - d_L)\, v(L) - u'(0) \underbrace{v(0)}_{=0} - \int_0^L \underbrace{(u''(x) + f(x))}_{=0}\, v(x)\, dx = (u'(L) - d_L)\, v(L)$$

for *any* $v \in \mathcal{V}$. In this case we can choose any $v \in \mathcal{V}$ that satisfies $v(L) \neq 0$. This implies that we need $u'(L) - d_L = 0$. So, the variational equation implies that u needs to satisfy (1.12c).

So far we proved the left-to-right implication in (1.82), that is, if $F(u, v) = 0$, then

$$0 = EL(u, x) = u''(x) + f(x) \qquad x \in (0, L) \tag{1.85a}$$

$$0 = EL(u, L) = u'(L) - d_L \tag{1.85b}$$

These are precisely (1.12a) and (1.12c).

Notice now that if u satisfies (1.85), then it satisfies (1.83) for any $v \in \mathcal{V}$, as it follows from replacing in (1.83). That is, we concluded the right-to-left implication in (1.82). Hence,

$$EL(u, x) = 0 \quad \forall\, x \in \omega = (0, L] \iff F(u, v) = 0 \quad \forall\, v \in \mathcal{V}.$$

Notice that ω does not include $x = 0$. The variational equation does not require u to satisfy the condition $u(0) = g_0$, since regardless of the value of $u(0)$, u would satisfy the variational equation as long as it satisfies (1.85). Boundary condition $u(0) = g_0$ needs to be explicitly required from u, it does not arise from satisfying the variational equation, so it is an essential boundary condition.

1.3.3.1 General Steps to Obtain the Euler-Lagrange Equations

We can summarize the general steps to find the Euler-Lagrange equations next, illustrating them with variational equation (1.9a), namely,

$$\int_\Omega \left[k(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)\right] dx \\ - k(L)d_L v(L) = \int_\Omega f(x)v(x)\, dx. \tag{1.86}$$

for any $v \in \mathcal{V}$, where

$$\mathcal{V} = \{w\colon [0, L] \to \mathbb{R} \text{ smooth } \mid w(0) = 0\}.$$

We proceed as follows:

1. **Integrate the variational equation by parts to eliminate all derivatives from the test function.** In this case, we want to eliminate all derivatives that appear on v, so we integrate by parts as many times as needed to do that. For our example in (1.9a),

$$\begin{aligned} 0 &= \int_0^L k(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x) - f(x)v(x)\,dx \\ &\quad - k(L)d_L v(L) \\ &= \int_0^L -k(x)u''(x)v(x) + b(x)u'(x)v(x) + c(x)u(x)v(x) - f(x)v(x)\,dx \\ &\quad + (k(L)u'(L) - k(L)d_L)v(L) - k(0)u'(0)v(0) \end{aligned}$$

2. **Group terms with v at the same location, and use conditions in $\mathcal{V}$.** A number of boundary terms will appear as a result of the integration by parts. Since functions in $\mathcal{V}$ often need to satisfy conditions at the boundary, use such conditions at this point. Then, gather all terms that involve the same values of the test function v, or its derivatives, in the domain and at the boundary. For our example, $v(0) = 0$, and we collect the terms containing $v(L)$ and $v(x)$:

$$\begin{aligned} 0 = \int_0^L \left[-k(x)u''(x) + b(x)u'(x) + c(x)u(x) - f(x)\right] v(x)\,dx \\ + k(L)(u'(L) - d_L)v(L). \end{aligned} \quad (1.87)$$

3. **Obtain the differential equation and potential boundary conditions.** We use the fact that the resulting expression should be valid for any $v \in \mathcal{V}$. Again, appealing to a simple and formal argument, this means that every term that multiplies a test function v at some point x should be equal to zero at that point, since the value of $v(x)$ can be chosen arbitrarily. For our example, this means that

$$0 = -k(x)u''(x) + b(x)u'(x) + c(x)u(x) - f(x) \qquad x \in (0, L) \quad (1.88a)$$

$$0 = k(L)(u'(L) - d_L), \quad (1.88b)$$

since they multiply $v(x)$ for $x \in (0, L)$ and $v(L)$, respectively. Notice that if both equations in (1.88) are satisfied, then variational equation (1.86) holds for all $v \in \mathcal{V}$.

Example 1.59 Let's find the Euler-Lagrange equations for the variational equation in Nitsche's method, c.f. Example 1.31. In this case, the function u satisfies variational equation (1.31):

$$\begin{aligned} \int_0^L u'(x)v'(x)\,dx + u'(0)v(0) - u(0)v'(0) + \mu u(0)v(0) = \\ \int_0^L f(x)v(x)\,dx + d_L v(L) - g_0 v'(0) + \mu g_0 v(0) \end{aligned} \quad (1.89)$$

for all $v \in \mathcal{V} = \{v\colon [0,L] \to \mathbb{R} \text{ smooth}\}$.

To obtain the differential equation and boundary conditions it implies, we first eliminate the derivative on v by integrating by parts, to get (the integrated-by-parts first term on the left hand side is on the first line):

$$u'(L)v(L) - u'(0)v(0) - \int_0^L u''(x)v(x)\,dx + u'(0)v(0)$$
$$- u(0)v'(0) + \mu u(0)v(0) = \int_0^L f(x)v(x)\,dx + d_L v(L) - g_0 v'(0) + \mu g_0 v(0)$$

Of course, we cannot eliminate the derivative from the term that contains $v'(0)$, so we will keep it. Notice also that a term that appeared when integrating by parts cancel with an existing term of the variational equation (the ones in color). Collecting terms with v evaluated at the same points, we obtain

$$\int_0^L (u''(x) + f(x))v(x)\,dx = (u'(L) - d_L)v(L) + (g_0 - u(0))v'(0) + \mu(u(0) - g_0)v(0). \quad (1.90)$$

At this point we can consider subsets of test functions to reach conclusions from each term. For $v \in \mathcal{V}$ be such that $v(0) = v(L) = v'(0) = 0$, we conclude that

$$\int_0^L (u''(x) + f(x))v(x)\,dx = 0.$$

Therefore, using the same rationale as in the first example, we conclude that

$$u''(x) + f(x) = 0 \qquad x \in (0,L). \quad (1.91)$$

Therefore u needs to satisfy this differential equation to satisfy the variational equation. For such u then, the first term in (1.90) is identically zero for *all* $v \in \mathcal{V}$ (not only for those that satisfy $v(0) = v(L) = v'(0) = 0$), and hence

$$0 = (u'(L) - d_L)v(L) + (g_0 - u(0))v'(0) + \mu(u(0) - g_0)v(0)$$

for all $v \in \mathcal{V}$. If we select $v \in \mathcal{V}$ such that $v'(0) = v(0) = 0$ and $v(L) = 1$ (for example, $v(x) = x^2/L^2$, then we conclude that $(u'(L) - d_L) = 0$, or that $u'(L) = d_L$, the Neumann boundary condition. We can again use this information, and assert that for *any* $v \in \mathcal{V}$,

$$0 = (g_0 - u(0))v'(0) + \mu(u(0) - g_0)v(0).$$

Here we have options for how to test. We could test with $v \in \mathcal{V}$ such that $v(0) = 1$ and $v'(0) = 0$ (for example, $v(x) = 1 - x^2$). This leads us to conclude that the first term $(g_0 - u(0))v'(0)$ is equal to zero, and that from there the second term implies that $u(0) = g_0$. Alternatively, we could have selected

a function for which $v(0) = 0$ and $v'(0) = 1$ (for example, $v(x) = x^2/2$, and conclude that the second term is identically zero, and that the first term implies that $u(0) = g_0$.

At this point, we can list the Euler-Lagrange equations we obtained as

$$u''(x) + f(x) = 0 \qquad x \in (0, L) \tag{1.92a}$$
$$u(0) = g_0 \tag{1.92b}$$
$$u'(L) = d_L. \tag{1.92c}$$

If (1.92) holds, then (1.90) is satisfied for all $v \in \mathcal{V}$. Hence there are no other Euler-Lagrange equations, since if u satisfies these three conditions, we can guarantee that variational equation (1.89) holds for all functions in $\mathcal{V}$.

It is evident from here that both boundary conditions are natural boundary conditions in this case.

1.3.4 The Weak and the Strong Forms

Consider the following problem:

Problem 1.3 (A Weak Form for Problem 1.1)**.** *Let*

$$\mathcal{S} = \left\{w \colon \Omega \to \mathbb{R} \text{ smooth } \mid w(0) = g_0\right\}, \tag{1.93a}$$
$$\mathcal{V} = \left\{w \colon \Omega \to \mathbb{R} \text{ smooth } \mid w(0) = 0\right\}. \tag{1.93b}$$

Find $u \in \mathcal{S}$ such that for any functions $v \in \mathcal{V}$

$$\int_\Omega \left[k(x)u'(x)v'(x) + b(x)u(x)'v(x) + c(x)u(x)v(x)\right] dx \\ - k(L)d_L v(L) = \int_\Omega f(x)v(x)\, dx. \tag{1.94}$$

This is called a **weak form** for Problem 1.1. A solution of Problem 1.1 is called a **weak solution**. Conversely, Problem 1.1 is called the **strong form**, and its solution is called the **strong solution**. The affine space $\mathcal{S}$ is the trial space, and $\mathcal{V}$ is the test space.

A solution u of Problem 1.1 is a solution of Problem 1.3, since it satisfies variational equation (1.94). Conversely, a solution of Problem 1.3 is a solution of Problem 1.1, since:

- A weak solution satisfies variational equation (1.94), and therefore it satisfies its Euler-Lagrange equations. Based on the discussion in §1.3.3.1, the Euler-Lagrange equations are the differential equation (1.6a) and the Neumann boundary condition (1.6c), namely,

$$-\left(k(x)u'(x)\right)' + b(x)u'(x) + c(x)u(x) = f(x) \qquad \forall\, x \in \Omega$$
$$u'(L) = d_L$$

- A weak solution to belongs to $\mathcal{S}$, so it also satisfies the Dirichlet boundary condition (1.6b).

$$u(0) = g_0.$$

Notice that variational equation (1.94) involves only the first derivative of u, while the strong form requires the second derivative of u to be defined as well. So, the conditions for a function to satisfy variational equation (1.94) are *weaker* than for the strong form. This means that, potentially, a functions that does not have well-defined second derivative could satisfy the variational equation.

If we expanded the trial space $\mathcal{S}$ to include more functions, functions that may have a first derivative but not a continuous second derivative, we could potentially find weak solutions[6] for which the strong form does not make sense, since a discontinuous second derivative implies that the differential equation in the strong form may not be defined at some points.

These same smoothness requirements need to be considered when stating the equivalence between a variational equation and its Euler-Lagrange equations, as in §1.3.3. If the function u that satisfies variational equation $F(u, v) = 0$ for all $v \in \mathcal{V}$ is not smooth enough for the Euler-Lagrange equation $EL(u, x) = 0$ to be defined for all x in the domain, then we cannot talk about equivalence.

This discussion is included here only for the reader to be aware of potential caveats. In this first approach to the subject, and to avoid delving into more tools in mathematics[7], we do not believe this is necessary. In the community of finite element analysts or practitioners, the term *weak form* is sometimes used to refer to the variational equation.

We conclude this section by stating a type of abstract weak forms, which have precisely the same structure as the linear variational method in Problem 1.2:

Problem 1.4 (Abstract Weak Form)**.** *Let $\mathcal{W}$ be a vector space, $a\colon \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ be a bilinear form, and $\ell\colon \mathcal{W} \to \mathbb{R}$ be a linear functional. Let the trial space $\mathcal{S}$ be an affine subspace of $\mathcal{W}$, and let the test space $\mathcal{V} \subset \mathcal{W}$ be the direction of $\mathcal{S}$.*

$$\text{Find } u \in \mathcal{S} \text{ such that } a(u, v) = \ell(v) \text{ for all } v \in \mathcal{V}. \tag{1.95}$$

1.4 The Finite Element Method

By now we have learned about how to construct a variational equation from the differential equation and boundary conditions of a problem, and to formulate a variational method to obtain an approximation of the solution. This last step relied on the construction of a trial and a test space in which to seek the approximate solution. We next look at how Finite Elements provide a systematic way to construct such spaces.

A Finite Element Method (FEM) is obtained by seeking the solution of a variational method in trial and test spaces constructed with finite elements.

[6] If, for example, $f(x)$ is discontinuous.

[7] For example, weak derivatives and Hilbert spaces.

We will also describe *how* to compute the stiffness matrix and load vector for a finite element method. The way this computation is performed, called **assembly**, is a distinctive feature and a virtue of the Finite Element method, since it can be done very efficiently in a computer.

1.4.1 The Simplest C^0 Finite Element Space

We show a first example of the construction of a variational method with the simplest finite element space of continuous functions. We do this for the problem in Example 1.55, so that we can contrast the use of variational methods with and without finite element spaces. In this example, we seek an approximation to the problem of finding $u\colon [0,1] \to \mathbb{R}$ such that

$$-u''(x) = 1 \qquad x \in (0,1) \tag{1.96a}$$
$$u(0) = 2 \tag{1.96b}$$
$$u'(1) = 0, \tag{1.96c}$$

using the following variational equation that u satisfies,

$$\int_0^1 u'(x)\,v'(x)\,dx = \int_0^1 v(x)\,dx \tag{1.97}$$

for all $v \in \mathcal{V} = \{w\colon [0,1] \to \mathbb{R} \text{ smooth } \mid w(0) = 0\}$.
Steps:

1. **Build the mesh of the domain.** Let the domain of the problem be the interval $\Omega = [c,d]$. We partition the domain into $n_{\text{el}} \in \mathbb{N}$ intervals by selecting $\{x_i\}_{i=1,\ldots,n_{\text{el}}+1}$ such that

$$c = x_1 < \ldots < x_{n_{\text{el}}+1} = d. \tag{1.98}$$

Each point x_i is a **vertex**, and i is its **vertex number**. Interval $[x_i, x_{i+1}]$ is called *element* i, for $i = 1,\ldots,n_{\text{el}}$. Strictly speaking, this is the domain of the element, but it is common to refer to the domain of the element simply as "element." The collection of nodes and elements is the **mesh**; we shall give a more complete definition of the mesh when we look at 2D problems.

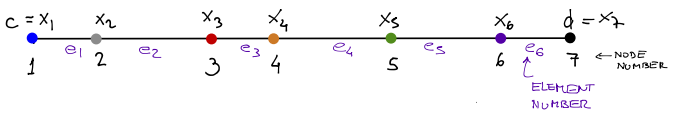


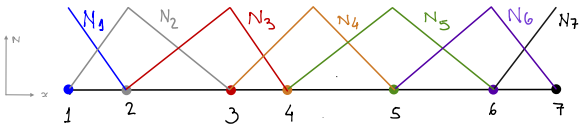
Figure 1.9

For our example, we choose a uniform mesh, so $c = 0, d = 1$, and $x_a = (a-1)/n_{\text{el}}$ for $i = 1,\ldots,n_{\text{el}}+1$.

2. **Build basis functions.** For this example, we build the so-called continuous "piecewise affine" elements, or hat functions. These functions have domain $[c,d]$, and for $a = 1,\ldots,n_{\text{el}}+1$ are defined as

$$N_a(x) = \begin{cases} 0 & x < x_{a-1} \\ \frac{x-x_{a-1}}{x_a-x_{a-1}} & x_{a-1} \le x < x_a \\ 1 & x = x_a \\ \frac{x_{a+1}-x}{x_{a+1}-x_a} & x_a < x \le x_{a+1} \\ 0 & x_{a+1} < x \end{cases} \tag{1.99}$$
$$= \max\left[0, \min\left(\frac{x-x_{a-1}}{x_a-x_{a-1}}, \frac{x-x_{a+1}}{x_a-x_{a+1}}\right)\right].$$

Notice that when $a = 1$, $x \in [c,d]$ implies that we only have the case $x \ge x_a$ [8]. Similarly, when $a = n_{\text{el}}+1$, $x \in [c,d]$ implies that we only have the case $x \le x_a$ [9]. These functions are plotted below.



The Finite Element space is $\mathscr{W}_h = \text{span}(\{N_1,\ldots,N_{n_{\text{el}}+1}\})$, so for this example, $m = n_{\text{el}}+1$. Some properties that will be generalized later to many other shape functions:

(a) You can check that all of the add up to 1, i.e., $\sum_{a=1}^{n_{\text{el}}+1} N_a(x) = 1$ for $x \in [c,d]$.

A simple way to see this, is to notice that in each element e, the two non-zero functions N_e and N_{e+1} are affine, and hence their sum is affine. But $N_e + N_{e+1}$ is equal to 1 at $x = x_e$ and $x = x_{e+1}$, and hence their sum is the only affine function that is equal to 1 at both locations: this is the constant function equal to 1.

(b) Notice that $N_b(x_a) = \delta_{ba}$. This is a particular version of a more general property we will see later in the class, and it has the following neat consequence. A function $w_h \in \mathscr{W}_h$ can be written as $w_h = w_1 N_1 + w_2 N_2 + \ldots + w_{n_{\text{el}}+1} N_{n_{\text{el}}+1}$, where $w_1,\ldots,w_{n_{\text{el}}+1}$ are the components of w_h in the basis. At the same time, $w_a = w_h(x_a)$ for $a = 1,\ldots,n_{\text{el}}+1$, that is, the component w_a is the value of the function w_h at x_a. This follows because

$$w_h(x_a) = w_1 \underbrace{N_1(x_a)}_{=\delta_{1a}} + \ldots + w_b \underbrace{N_b(x_a)}_{=\delta_{ba}} + \ldots + w_{n_{\text{el}}+1} \underbrace{N(x_a)}_{=\delta_{(n_{\text{el}}+1)a}} = w_a,$$

[8] We do not know what x_{a-1} is in this case, nor do we need it.
[9] We do not know what x_{a+1} is in this case, but again, we do not need it.

so it is a special property of the basis we chose for $\mathcal{W}_h$.

Had we chosen the basis $\{N_1 + N_2, N_2, \ldots, N_{n_{\text{el}}+1}\}$, for example, then $w_h(x_2) = w_1 + w_2$, and in this case w_2 does not necessarily coincide with the value of w_h at x_2.

To indicate that the degrees of freedom $\{w_1, \ldots, w_{n_{\text{el}}+1}\}$ of the function w_h in the basis $\{N_1, \ldots, N_{n_{\text{el}}+1}\}$ are precisely the values of w_h at each vertex x_a, we say that there is a **node** of the finite element space at each vertex of this mesh, and graphically depict it with a filled disk at the vertex; see Fig. 1.9.

(c) Notice that $N_a(x) \neq 0$ only in a small part of the domain. This is normally referred to by saying that the basis functions have "compact support." In the finite element context, this (generally) means that basis functions are non-zero in at most one element and its neighbors.

In this case we defined the space $\mathcal{W}_h$ as the span of a set of basis functions. Alternatively, it could have been defined as

$$\mathcal{W}_h = \{w_h \colon [c,d] \to \mathbb{R} \text{ continuous } \mid w_h \text{ is affine on each element}\}. \quad (1.100)$$

This space would often be referred to as the "space of piecewise affine functions over $[c,d]$," with the tacit understanding that functions would be affine over each element. Both definitions are equivalent; it is simple to see that functions in $\text{span}(\{N_1, \ldots, N_{n_{\text{el}}+1}\})$ are piecewise affine, and that any piecewise affine function can be expressed as a linear combination of functions in the basis $\{N_1, \ldots, N_{n_{\text{el}}+1}\}$.

3. **Build $\mathcal{V}_h$ and $\mathcal{S}_h$.** Collect essential boundary conditions and impose them on functions in $\mathcal{W}_h$ to obtain $\mathcal{S}_h$. The space $\mathcal{V}_h$ follows as the direction of $\mathcal{S}_h$. In our example,

$$\begin{aligned}\mathcal{S}_h &= \{u_h \in \mathcal{W}_h \mid u_h(0) = 2\},\\ \mathcal{V}_h &= \{v_h \in \mathcal{W}_h \mid v_h(0) = 0\}.\end{aligned}$$

To find a basis for $\mathcal{V}_h$, notice that for any $v_h = v_1 N_1 + \ldots, + v_{n_{\text{el}}+1} N_{n_{\text{el}}+1} \in \mathcal{W}_h$, $v_h(0) = 0$ if and only if $v_1 = 0$. Similarly, any function $u_h \in \mathcal{W}_h$ satisfies that $u_h(0) = 2$ if and only if $u_1 = 2$. Then, in terms of the basis functions in $\mathcal{W}_h$, these spaces can be described by

$$\begin{aligned}\mathcal{V}_h &= \{v_2 N_2 + \ldots + v_{n_{\text{el}}+1} N_{n_{\text{el}}+1} \mid v_2, \ldots, v_{n_{\text{el}}+1} \in \mathbb{R}\}\\ &= \text{span}\left(\{N_2, \ldots, N_{n_{\text{el}}+1}\}\right).\\ \mathcal{S}_h &= \{u_h \in \mathcal{W}_h \mid u_1 = 2\}\\ &= \{2N_1 + v_h \mid v_h \in \mathcal{V}_h\}.\end{aligned}$$

The index set η_a that identifies the basis functions for $\mathcal{V}_h$, and its complement η_g, are

$$\begin{aligned}\eta_a &= \{2, \ldots, n_{\text{el}} + 1\}\\ \eta_g &= \{1\}.\end{aligned}$$

Finally, we need to identify the components $\overline{u}_a$ for $a \in \eta_g$ to impose the fact that $u_h \in \mathcal{S}_h$. In this case, based on the description of $\mathcal{S}_h$ above, it is $\overline{u}_1 = 2$.

As an exercise, we can also identify a function $\overline{u}_h \in \mathcal{S}_h$. For example, we can choose $\overline{u}_h = 2N_1$, which gives $\overline{u}_1 = 2$. Alternatively, we can set $\overline{u}_h = 2 = 2\sum_{a\in\eta} N_a$, which is also in $\mathcal{S}_h$ because it is in $\mathcal{W}_h$ and $\overline{u}_h(0) = 2$.

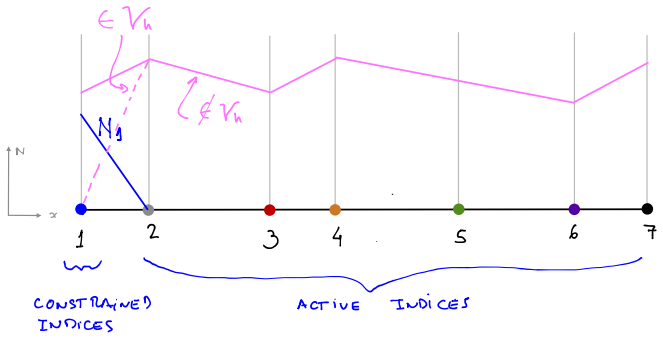


Figure 1.10 If $w_1 \neq 0$, then $w_h \notin \mathcal{V}_h$.

Notice that the constraint imposed on u_1 stems from the essential boundary condition (EBC). Should more EBC be present, more constraints should be imposed on functions in $\mathcal{W}_h$ to belong to $\mathcal{V}_h$ and $\mathcal{S}_h$. The EBC will also determine components of $\overline{u}_a$ for $a \in \eta_g$.

Not every constraint can be imposed by simply selecting a subset of a set of basis functions for $\mathcal{W}_h$, as we have assumed so far. For example, had we chosen the basis $\{N_1, N_1 + N_2, N_3, \ldots, N_{n_{el}+1}\}$ for $\mathcal{W}_h$, then $w_h(0) = w_1 + w_2$, and the condition for $w_h \in \mathcal{W}_h$ to belong to $\mathcal{V}_h$ is to have $w_1 + w_2 = 0$, see Fig. 1.11. For example, the function $w_h = N_1 + (N_2 - N_1)$ has $w_1 = 1$ and $w_2 = 1$ and is in $\mathcal{V}_h$. Therefore, just setting either w_1 or w_2 (or both) to zero does not lead to a basis for $\mathcal{V}_h$. It is not possible in this case to extract a subset of $\{N_1, N_1 + N_2, N_3, \ldots, N_{n_{el}+1}\}$ to serve as a basis for $\mathcal{V}_h$[10]. As a result, it is not possible to define η_a or η_g.

In the finite element method, this type of situations need a different treatment (e.g., with Lagrange multipliers or Nitsche's method), and the most commonly used finite element bases are constructed so that essential boundary conditions can be imposed by setting the values of some components, such as $v_1 = 0$ here. In other words, in the finite element method it is common for the basis for $\mathcal{V}_h$ to be a subset of the basis for $\mathcal{W}_h$. This is going to be the case for the examples we will see.

Summarizing, in this step we:

(a) Identify active and constrained index sets, η_a and η_g (this assumes that the basis for $\mathcal{V}_h$ is a subset of the basis for $\mathcal{W}_h$).

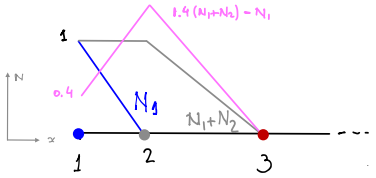


Figure 1.11 The function $w_h = 1.4(N_1 + N_2) - N_1$ has $w_1 = -1$ and $w_2 = 1.4$ in the basis $\{N_1, N_1 + N_2, \ldots\}$, and $w_h(0) = w_1 + w_2 = 0.4 \neq w_1$. Therefore, to impose the EBC at $x = 0$ in this basis it is not enough to solely set the value of w_1.

[10] Of course, $\mathcal{V}_h$ has a basis, but it is not a subset of the chosen basis for $\mathcal{W}_h$.

(b) Build $\mathcal{V}_h = \operatorname{span}\left(\{N_a \in \{N_1, \dots, N_m\} \mid a \in \eta_a\}\right)$.

(c) Identify $\overline{u}_h \in \mathcal{W}_h$ so that $\mathcal{S}_h = \{v + \overline{u}_h \mid v \in \mathcal{V}_h\} = \left\{v + \sum_{a\in\eta_g} \overline{u}_a N_a \mid v \in \mathcal{V}_h\right\}$.

4. **Compute K and F.** We proceed as we did earlier and compute the stiffness matrix and load vector. We compute its entries according to (1.78).

   For our example, we can set $h = 1/n_{\text{el}}$ with $n_{\text{el}} = 5$ and hence $m = 6$,

$$\ell(N_a) = \int_0^1 1 \cdot N_a(x)\, dx = \begin{cases} \dfrac{h}{2} & a \in \{1, m\} \\ h & a \in \{2, \dots, m-1\}. \end{cases}$$

$$a(N_b, N_a) = \int_0^1 N_b'(x) N_a'(x)\, dx = \begin{cases} 0 & |a-b| > 1 \\ -\dfrac{1}{h} & |a-b| = 1 \\ \dfrac{2}{h} & a = b \in \{2, \dots, m-1\}. \\ \dfrac{1}{h} & a = b \in \{1, m\}. \end{cases}$$

   The only index in η_g is 1. Therefore, according to (1.78c),

$$\begin{aligned} K_{21} &= a(N_1, N_2) = -\frac{1}{h}, \\ K_{12} &= \delta_{12} = 0, \\ K_{11} &= \delta_{11} = 1, \\ K_{22} &= a(N_2, N_2) = \frac{2}{h}, \\ K_{23} &= a(N_3, N_2) = -\frac{1}{h}, \\ K_{24} &= a(N_4, N_2) = 0, \\ K_{66} &= a(N_6, N_6) = \frac{1}{h}, \\ F_1 &= \overline{u}_1 = 2, \\ F_5 &= \ell(N_5) = h, \\ F_6 &= \ell(N_6) = \frac{h}{2}. \end{aligned}$$

   We have not replaced $h = 1/m = 1/5$ yet, for clarity. In this case, the stiffness matrix and load vector are, now replacing $h = 1/5$,

$$K = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -5 & 10 & -5 & 0 & 0 & 0 \\ 0 & -5 & 10 & -5 & 0 & 0 \\ 0 & 0 & -5 & 10 & -5 & 0 \\ 0 & 0 & 0 & -5 & 10 & -5 \\ 0 & 0 & 0 & 0 & -5 & 5 \end{bmatrix} \qquad F = \begin{bmatrix} 2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.1 \end{bmatrix}. \tag{1.101}$$

5. **Solve and Compute the Finite Element Solution.** We now solve the system $KU = F$, and then build the finite element solution as $u_h(x) = \sum_{a=1}^{m} u_a N_a(x)$.

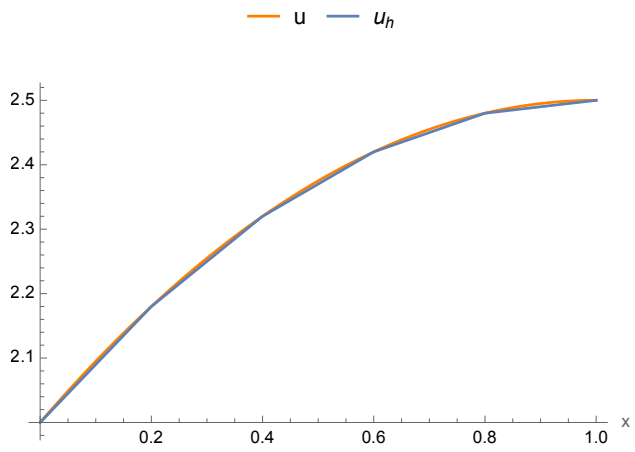
   For our example,

$$U = \begin{bmatrix} 2 \\ 2.18 \\ 2.32 \\ 2.42 \\ 2.48 \\ 2.5 \end{bmatrix}$$

   and hence

$$u_h(x) = 2N_1(x) + 2.18N_2(x) + 2.32N_3(x) + 2.42N_4(x) + 2.48N_5(x) + 2.5N_6(x).$$

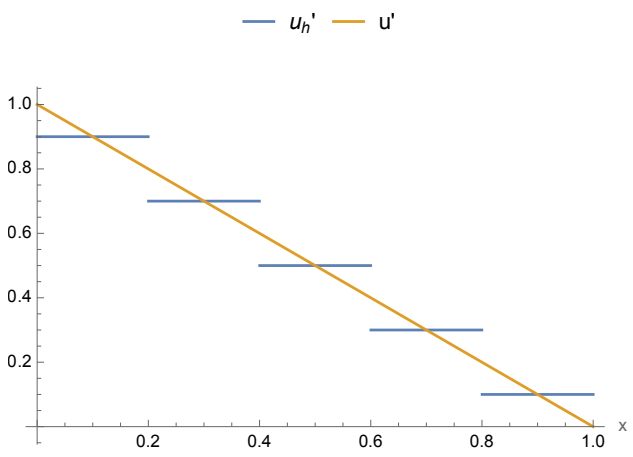
   This function is plotted below, together with the exact solution $u(x) = 2 + x - x^2/2$.



   The derivative of u_h follows as

$$u_h'(x) = 2N_1'(x) + 2.18N_2'(x) + 2.32N_3'(x) + 2.42N_4'(x) + 2.48N_5'(x) + 2.5N_6'(x).$$

   The derivative of the finite element approximation is a piecewise constant function, a fact reflected in its graph, shown below.



In staring at this graph, you may come to the realization that the finite element approximation cannot even be evaluated as a candidate solution to the differential equation of the problem, (1.96a), since neither the first nor the second derivatives are defined at the vertices of the mesh. Yet, we did obtain a good approximation of the exact solution in this way. This observation highlights why finite element methods are constructed from a variational equation of the problem: (we will see that) it is simpler to construct functions if milder smoothness requirements are imposed.

We conclude by showing the stiffness matrix of this problem for a constant h and any m, in the case in which $\mathcal{V}_h = \mathcal{W}_h$ (no EBC). This is a matrix also found in finite differences in the same problem, and hence it is a commonly found matrix in elementary numerical analysis textbooks. According to step 4, the matrix is

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

1.4.1.1 About Consistency.

In creating the variational method for this example we selected a space $\mathcal{V}_h$ that contains functions that are not smooth (the first derivative is generally discontinuous). Therefore, $\mathcal{V}_h$ is *not* a subset of the test space $\mathcal{V}$ in the variational equation we started from, (1.97). Since $\mathcal{V}_h \not\subseteq \mathcal{V}$, we cannot immediately tell that the method is consistent. Instead, we have to check its consistency, namely, we need to check if $F(u, v_h) = 0$ for all $v_h \in \mathcal{V}_h$

This is true in this case, and the method is consistent. To see this, we can use the integration by parts formula for piecewise smooth functions (1.45) in Theorem 1.2. Since any $v_h \in \mathcal{V}_h$ is a continuous function by construction, then the

same integration-by-parts formula used for smooth functions holds. We proceed as we do when finding the Euler-Lagrange equations, we integrate by parts the left hand side of (1.97) to eliminate derivatives over the test function v_h:

$$\begin{aligned} F(u, v_h) &= \int_0^1 u'(x) v_h'(x)\, dx - \int_0^1 v_h(x)\, dx = \\ &= \underbrace{u'(1) v_h(1)}_{=0, \text{ due to (1.96c)}} - \underbrace{u'(0) v_h(0)}_{=0, v_h \in \mathcal{V}_h} - \int_0^1 u''(x) v_h(x)\, dx - \int_0^1 v_h(x) dx \\ &= -\int_0^1 \underbrace{(u''(x) + 1)}_{=0, \text{ due to (1.96a)}} v_h(x)\, dx \\ &= 0. \end{aligned}$$

The test space $\mathcal{V}$ is commonly defined so that it already includes continuous finite element functions. This is convenient but not needed.

Therefore, we can state that

$$F(u, v) = 0 \qquad \forall v \in \mathcal{V} + \mathcal{V}_h,$$

where $\mathcal{V} + \mathcal{V}_h = \{w = v + v_h \mid v \in \mathcal{V}, v_h \in \mathcal{V}_h\}$.

The key ingredients to prove consistency here were that functions are smooth inside each element, and continuous across them, since in this case the integration by parts formula 1.2 holds.

1.4.2 What is a Finite Element?

We proceed now to describe how we construct finite element spaces beyond the span of hat functions that we introduced in §1.4.1. The construction of finite element spaces is done in two steps: (1) definition of vector spaces of functions over each element, and (2) adding functions defined over different elements to form functions whose domain is the entire interval $[c, d]$. We describe the first step next, and the second step in the next section, §1.4.3.

We begin by introducing the definition of a finite element.

Definition 1.11 (Finite Element)**.** *A finite element is a pair $e = (\Omega_e, \mathcal{N}^e)$ of an element domain Ω_e and a finite set of basis functions $\mathcal{N}^e = \{N_1^e, \ldots, N_k^e\}$ defined over Ω_e.*

Given a finite element $e = (\Omega_e, \mathcal{N}_e)$ with element domain Ω_e and a set $\mathcal{N}^e = \{N_1^e, \ldots, N_k^e\}$ of linearly independent functions $N_i^e : \Omega_e \to \mathbb{R}$, the space of functions $\mathscr{P}^e$ over Ω_e is defined as

$$\mathscr{P}^e = \text{span}\{N_1^e, \ldots, N_k^e\} \tag{1.102}$$

for $k \geq 1$, and it is called the **element space**. The set $\mathcal{N}^e$ is a basis for $\mathscr{P}^e$. Functions in $\mathcal{N}^e$ are known as **shape functions**. The number of shape functions k, or dimension of $\mathscr{P}^e$, is the **number of degrees of freedom** of the element. The **degrees of freedom** of the element are the components $\{\phi_1^e, \ldots, \phi_k^e\}$ of functions in this basis. Each one of the components ϕ_i^e, $i = 1, \ldots, k$, is a variable that can

take any real value, and hence the k-tuple $(\phi_1^e,\ldots,\phi_k^e)$ can take any value in $\mathbb{R}^k$. For each such value, a unique function $f^e \in \mathscr{P}^e$, $f^e\colon \Omega_e \to \mathbb{R}$, is defined through

$$f^e(x) = \phi_1^e N_1^e(x) + \ldots + \phi_k^e N_k^e(x) = \sum_{a=1}^{k} \phi_a N_a^e(x). \tag{1.103}$$

The symbol e will be used interchangeably to denote an element or an **element index**, often a natural number, given that the index is another way to identify what element we are referring to. It is also common to use the word element in lieu of element domain; for example, wording such as ... *integrating over an element*..., as a way to say integrating over Ω_e.

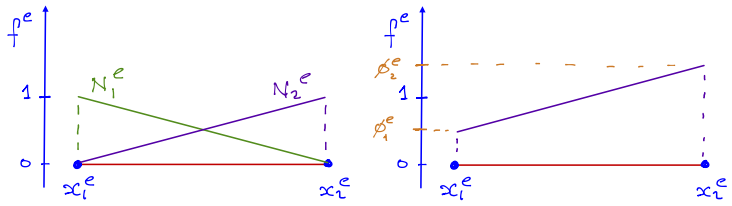
For the following examples we will consider a generic element with element domain $\Omega_e = [x_1^e, x_2^e]$, with vertex 1 at x_1^e and vertex 2 at x_2^e.

Examples:

1.60 P_1**-element.** One of the simplest element spaces is generated by the basis functions

$$\begin{aligned} N_1^e(x) &= \frac{x - x_2^e}{x_1^e - x_2^e}, \\ N_2^e(x) &= \frac{x - x_1^e}{x_2^e - x_1^e}, \end{aligned} \tag{1.104}$$

which satisfy that $N_a^e(x_b^e) = \delta_{ab}$.



To see that the two are linearly independent, let

$$f^e(x) = \phi_1^e \frac{x - x_2^e}{x_1^e - x_2^e} + \phi_2^e \frac{x - x_1^e}{x_2^e - x_1^e}$$

and assume that $f^e(x) = 0$ for all $x \in \Omega_e$. In particular, $f^e(x_1^e) = \phi_1^e = 0$, and similarly, $f^e(x_2^e) = \phi_2^e = 0$. Therefore, this is a set of linearly independent functions.

The space $\mathscr{P}^e$ has 2 degrees of freedom, it is the space $\mathbb{P}_1(\Omega_e)$ of all polynomials of degree 1 or less over Ω_e. To see this, notice that $N_1^e(x) + N_2^e(x) = 1$ for all x, and $x_1^e N_1^e(x) + x_2^e N_2^e(x) = x$, so $\{1, x\} \in \mathscr{P}^e$. The degrees of freedom here are the values of f^e at x_1^e and x_2^e; this is the interpretation of ϕ_1^e and ϕ_2^e. Thus, we say that this element has a node at x_1^e and a node at x_2^e, and indicated them with a filled disc as follows

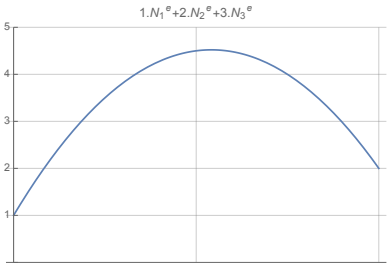


Figure 1.12 A function in the P_1-element+bubble.

1.61 A variation of the P_1-element has the basis

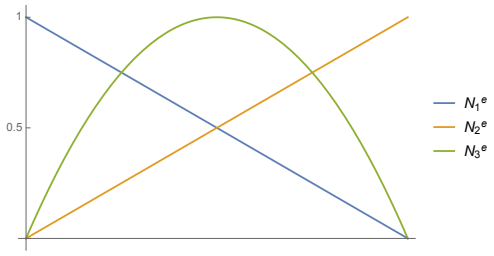
$$\begin{aligned} N_1^e(x) &= 1 \\ N_2^e(x) &= x. \end{aligned} \tag{1.105}$$

The space $\operatorname{span}\{N_1^e, N_2^e\}$ is still $\mathbb{P}_1(\Omega_e)$. However, the degrees of freedom in this case do not always lend themselves to be interpreted as pointwise values of the function $f^e = \phi_1^e\, 1 + \phi_2^e x$ somewhere in the element. There is no standard graphical depiction of this element.

1.62 **P_1-element+bubble.** Next, consider the basis functions

$$\begin{aligned} N_1^e(x) &= \frac{x - x_2^e}{x_1^e - x_2^e}, \\ N_2^e(x) &= \frac{x - x_1^e}{x_2^e - x_1^e}, \\ N_3^e(x) &= 4 N_1^e(x) N_2^e(x). \end{aligned} \tag{1.106}$$

Their plot is



It is simple to check that this is a set of linearly independent functions. A function $f^e \in \mathscr{P}^e$ has the form

$$f^e(x) = \phi_1^e \frac{x - x_2^e}{x_1^e - x_2^e} + \phi_2^e \frac{x - x_1^e}{x_2^e - x_1^e} + \phi_3^e 4 \frac{(x - x_1^e)(x_2^e - x)}{(x_1^e - x_2^e)^2},$$

and one example of such function is shown in Fig. 1.12.

The space $\mathscr{P}^e$ has 3 degrees of freedom, and it is the space $\mathbb{P}_2(\Omega_e)$ of all polynomials of degree 2 or less over Ω_e. To see this, notice that $1, x \in \mathscr{P}^e$ from example 1.60, and that $x^2 = (x_2^e)^2 N_2^e(x) + (x_1^e)^2 N_1^e(x) - (x_2^e - x_1^e)^2/4\, N_3^e(x)$, so $x^2 \in \mathscr{P}^e$.

The degrees of freedom of this element do not all have a simple interpretation: ϕ_1^e and ϕ_2^e are the values of f^e at x_1^e and x_2^e, but ϕ_3^e lacks one. The name *bubble* comes from the shape of N_3^e, which is zero at the two boundaries of the element.

What is a node? In general, whenever a degree of freedom of an element is the value of the function f^e or one of its derivatives at a location $\bar{x}$, we say that the element has a *node* at $\bar{x}$. When the degree of freedom is the value of the function, we depict it with a filled disk at $\bar{x}$. The symbol to depict the value of a derivative as a degree of freedom will be introduced later.

As a counterexample, degree of freedom ϕ_3^e in the P_1-element+bubble (Example 1.62) does not always correspond to the value of a function in the space at the midpoint between x_1^e and x_2^e, so such degree of freedom cannot be indicated by a node, c.f. Fig. 1.12.

The pictorial depiction of nodes in an element is a way to graphically indicate the degrees of freedom of an element, and it is commonly used in the finite element literature.

It is important to retain a strict distinction between the vertices of an element, which are used to define the geometry of the element domain, and the degrees of freedom indicated by the nodes, which are used to define functions over the element domain.

Examples:

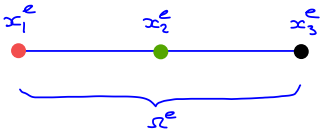
1.63 **P_0-element.** The simplest element space is that of a single constant function over the domain of the element Ω_e, or in a fancy way, a polynomial in $\mathbb{P}_0(\Omega_e)$. A single basis function is needed,

$$N_1^e(x) = 1, \tag{1.107}$$

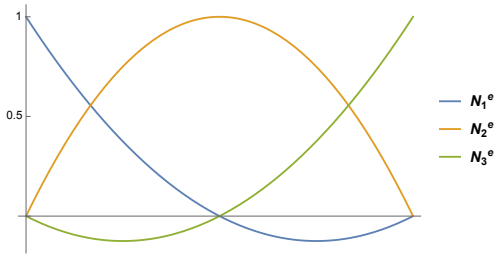
so the space $\mathscr{P}^e$ has one degree of freedom.

It can be represented with a node at the center of the element, or elsewhere.

1.64 **P_2-element.** The second most common element has the following basis functions over an element domain $\Omega_e = [x_1^e, x_3^e]$, with $x_2^e = (x_1^e + x_3^e)/2$,



$$\begin{aligned} N_1^e(x) &= \frac{(x-x_2^e)(x-x_3^e)}{(x_1^e-x_2^e)(x_1^e-x_3^e)}, \\ N_2^e(x) &= \frac{(x-x_1^e)(x-x_3^e)}{(x_2^e-x_1^e)(x_2^e-x_3^e)}, \\ N_3^e(x) &= \frac{(x-x_1^e)(x-x_2^e)}{(x_3^e-x_1^e)(x_3^e-x_2^e)}. \end{aligned} \tag{1.108}$$

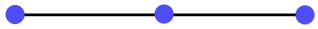


This is a linearly independent set of functions, a fact that follows from similar arguments to those for the P_1-element.

The space $\mathscr{P}^e$ has 3 degrees of freedom, it is also the space $\mathbb{P}_2(\Omega_e)$ of all polynomials of degree 2 or less over Ω_e. This is a set of three linearly independent quadratic polynomials, precisely the dimension of $\mathbb{P}_2(\Omega_e)$, and hence they need to span $\mathbb{P}_2(\Omega_e)$.

In this element we have that $N_1^e(x)+N_2^e(x)+N_3^e(x)=1$ for any $x\in\Omega_e$. To see this, let $f^e(x)=N_1^e(x)+N_2^e(x)+N_3^e(x)$, and notice that $f^e(x)=1$ for $x\in\{x_1^e,x_2^e,x_3^e\}$ and that $f^e\in\mathscr{P}^e$. Thus, $f^e(x)-1$ is a quadratic polynomial that is equal to 0 at these three points. We conclude then that $f^e(x)-1=0$ for all $x\in\Omega_e$. You can also check this by simply adding the three expressions in (1.108).

Since the basis functions satisfy that $N_a^e(x_b^e)=\delta_{ab}$, the degrees of freedom in the space are the values of a function $f^e\in\mathscr{P}^e$ at x_1^e, x_2^3 and x_3^e. Because these three spatial locations have the value of a function therein as a degree of freedom, the element has three nodes, each one represented with a filled disc, one at each vertex and one at x_3^e, to wit:



1.65 **P_k-element, for $k=1,\ldots$.** The Lagrange P_k-elements, often known simply as the P_k-elements, are a generalization of the P_0, P_1 and P_2 elements to any positive integer k. To simplify notation, we will denote the position of the vertices of the element by $z_1<z_2$, so that $\Omega_e=[z_1,z_2]$. Additionally, we introduce $k+1$ nodes at locations

$$x_a^e=z_1+(a-1)\frac{(z_2-z_1)}{k}$$

for $a=1,\ldots,k+1$. The basis functions for this element are

$$N_a^e(x)=\frac{\Pi_{b=1,b\neq a}^{k+1}(x-x_b^e)}{\Pi_{b=1,b\neq a}^{k+1}(x_a^e-x_b^e)} \tag{1.109}$$

for $a=1,\ldots,k+1$. Each of these functions is a polynomial of degree k, and as will see next, they form a linearly independent set of $k+1$ functions. Therefore, $\mathscr{P}^e=\text{span}(N_1^e,\ldots,N_{k+1}^e)=\mathbb{P}_k(\Omega_e)$, or the set of all polynomials of degree less or equal than k over Ω_e, since the number

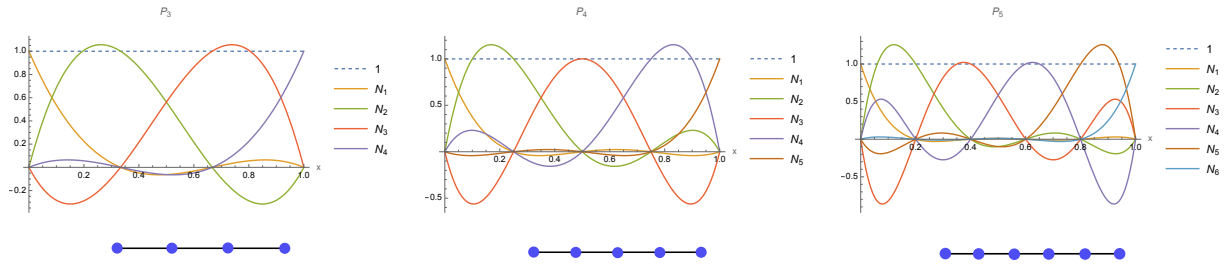


Figure 1.13 Shape functions for elements P_3, P_4 and P_5 over $\Omega_e = [0,1]$, together with the constant function $f(x) = 1$ for comparison. The graphical depiction of each element is shown as well.

of linearly independent vectors functions is equal to the dimension of $\mathbb{P}_k(\Omega_e)$. The plots of these basis functions for $k = 3,4,5$ are shown in Fig. 1.13.

The first noteworthy feature of this set of functions is that

$$N_a^e(x_b^e) = \delta_{ab}, \tag{1.110}$$

so that the degrees of freedom of this element are the values of a function at $\{x_1^e, \ldots, x_{k+1}^e\}$. Therefore, this element has nodes at these locations. To see (1.110), notice that if $a \neq b$, then x_b^e is a zero of the numerator of N_a^e. Instead, if $a = b$, then the numerator and denominator of (1.110) are equal, and hence $N_a^e(x_a^e) = 1$.

To see that this is a basis, consider $(\phi_1^e, \ldots, \phi_{k+1}^e) \in \mathbb{R}^{k+1}$ such that

$$f(x) = \phi_1^e N_1^e(x) + \ldots + \phi_{k+1}^e N_{k+1}^e(x) = 0 \qquad \forall x \in \Omega_e.$$

Then, for any $a = 1, \ldots, k+1$, $f(x_a^e) = \phi_a^e N_a^e(x_a^e)$, because $N_b^e(x_a^e) = 0$ for $a \neq b$, from where $\phi_a^e = 0$. It then follows that this is a basis.

The final interesting property of this basis is that if $f \in \mathbb{P}_k(\Omega_e)$, then

$$f(x) = f(x_1^e) N_1^e(x) + \ldots + f(x_{k+1}^e) N_{k+1}^e(x) \qquad \forall x \in \Omega_e. \tag{1.111}$$

In particular, if $f(x) = 1$, then $N_1^e(x) + \ldots + N_{k+1}^e(x) = 1$ for all $x \in \Omega_e$. To prove (1.111), let $g(x) = f(x_1^e) N_1^e(x) + \ldots + f(x_{k+1}^e) N_{k+1}^e(x) - f(x)$. Notice then that $g(x_a^e) = 0$ for $a = 1, \ldots, k+1$, and that $g(x)$ is a polynomial of degree less or equal than k that is equal to zero at $k+1$ distinct points. This can only happen if $g(x) = 0$ for all $x \in \Omega_e$, from where (1.111) follows.

Elements in which all the degrees of freedom are values of the function at predefined locations in the element are called **Lagrange elements**. For example, the P_k-element is a Lagrange element, while the P_1-element+bubble is not.

1.4.3 Construction of Finite Element Spaces

Once we define a mesh over the interval Ω and element spaces on each element, we have what is called a **finite element mesh**. A vector space $\mathcal{W}_h$ of functions over the interval Ω can be constructed by defining a basis for it using the shape functions in each finite element.

What type of basis functions for $\mathcal{W}_h$ can we construct with the shape functions in each element? Let's look at some examples (we do not specify the entire basis yet). To this end, we consider a mesh of three P_1-elements in Fig. 1.14.

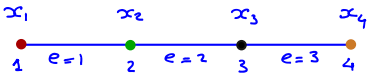
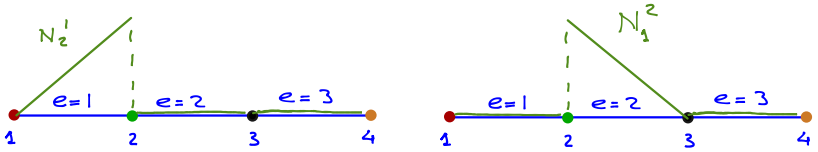


Figure 1.14

Examples:

1.66 Function N_2^1 could be a basis function, if we define it as equal to zero for points outside element $e = 1$. The function is discontinuous at x_2, so has two one-sided limits, $\lim_{x\to x_2^-} N_2^1(x) = 1$, and $\lim_{x\to x_2^+} N_2^1(x) = 0$. Similarly, N_2^1 could be a basis function, if we define it as equal to zero outside element $e = 2$, and it is discontinuous at x_2 as well. These two functions are sketched next:



1.67 The hat function N_2 can be constructed as the sum of N_1^2 and N_2^1, $N_2 = N_2^1 + N_1^2$, when each of them is defined as equal to zero outside elements $e = 1$ and $e = 2$, respectively, see Fig. 1.15. Because N_2^1 and N_1^2 are discontinuous at x_2, the value of N_2 at x_2 depends on what values each one of them takes at x_2. If we defined $N_2^1(x_2) = N_1^2(x_2) = 1$, then $N_2(x_2) = 2$, and N_2 would be discontinuous at x_2.

Instead, it is convenient to define

$$N_2(x) = N_2^1(x) + N_1^2(x) \qquad \text{for } x \neq x_2,$$

and define the value of N_2 at x_2 only if the two one-sided limits are the same. In this case they are, so

$$N_2(x_2) = \lim_{x\to x_2^-} N_2(x) = \lim_{x\to x_2^+} N_2(x) = \lim_{x\to x_2} N_2(x) = 1.$$

We will not consider cases in which the two one-sided limits are not equal. But if they were, the value of $N_2(x_2)$ need not be defined, and instead the method would use the values of the two one-sided limits. This is what is typically done in *Discontinuous Galerkin Methods*.

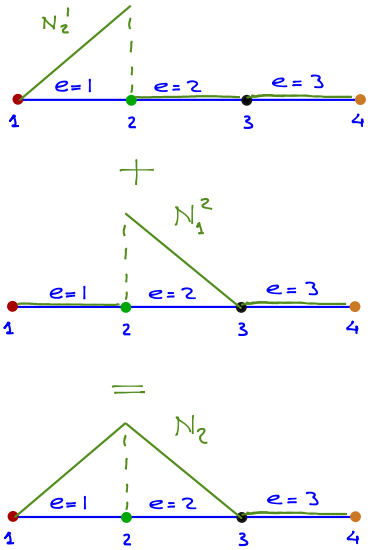


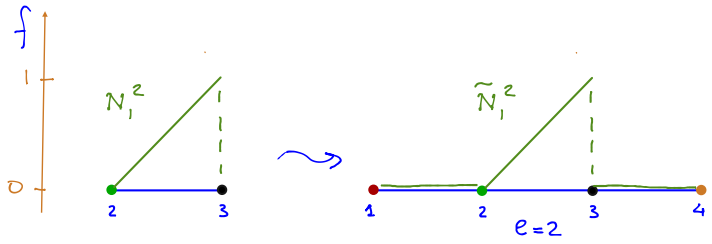
Figure 1.15

In general, to build a basis for $\mathcal{W}_h$, we proceed as follows:

1. **Extend Shape Functions by Zero.** For each element e in the mesh, we extend each basis function in the element by zero outside the element. More precisely, if we denote by $\tilde{N}_a^e \colon \Omega \to \mathbb{R}$ the extension-by-zero of the function $N_a^e \colon \Omega_e \to \mathbb{R}$, we can write

$$\tilde{N}_a^e(x) = \begin{cases} N_a^e(x) & x \in \Omega_e \\ 0 & x \notin \Omega_e. \end{cases} \tag{1.112}$$

See the example below.



In the following, we will not make an explicit distinction between $\tilde{N}_a^e$ and N_a^e, and simply use N_a^e for both.

2. **Define a Local-to-Global Map.** We will define *every* basis functions N_A for $\mathcal{W}_h$, with $A \in \{1, \ldots, m\}$, by *adding* shape functions from one or more elements, with the condition that

each shape function in an element contributes to exactly one basis function in $\mathcal{W}_h$.

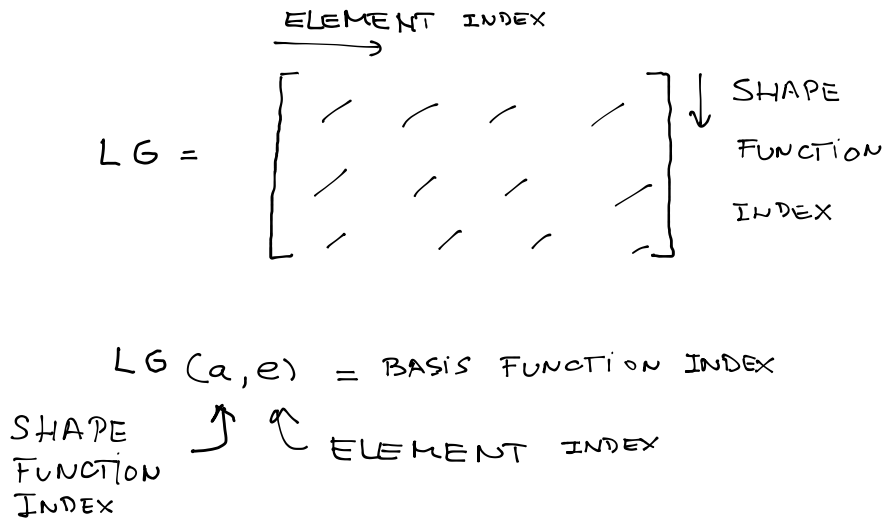
Then, $\mathcal{W}_h = \text{span}(N_1, \ldots, N_m)$ and its dimension is m. Symbolically, if N_A is the result of adding $r \geq 1$ shape functions, we can write

$$N_A = N_{a_1}^{e_1} + \ldots + N_{a_r}^{e_r},$$

and each shape function N_a^e appears in exactly one of such sums.

We specify what basis function N_A a shape function N_a^e contributes to through a **local-to-global map**.

In this class, the local-to-global map is indicated with an $k \times n_{\text{el}}$ matrix termed LG, so that $A = \text{LG}(a, e)$ is the entry in row a and column e in LG. Graphically:



The entry $\mathsf{LG}(a,e)$ in the matrix is a number in $\{1,\dots,m\}$ that defines that shape function N_a^e should be added when constructing basis function N_A. Alternatively, function N_A is obtained by adding all shape functions with index A in the matrix LG. Because each shape function contributes to exactly one basis function, every entry in the LG matrix is well-defined (there is no ambiguity on what value should go in entry of LG). The range of LG needs to be $\{1,\dots,m\}$, so that all basis functions in $\mathcal{W}_h$ are constructed in this way.

The name local-to-global map originates in the fact that it maps the indices of shape functions in each element, defined only locally over the domain of the element to form the element space, to indices of basis functions whose domain is the entire interval Ω to form the space $\mathcal{W}_h$.

3. **Add Shape Functions.** Because we are adding functions that are potentially discontinuous at the interfaces between neighboring elements, some care is needed in the definition of the basis functions for $\mathcal{W}_h$, as in Example 1.67.

   With the local-to-global map, we define the basis functions for $\mathcal{W}_h$. For any $A \in \{1,\dots,m\}$, let $N_A\colon \Omega \to \mathbb{R}$ be given by

$$N_A(x) = \sum_{\{(a,e)\,|\,\mathsf{LG}(a,e)=A\}} N_a^e(x), \tag{1.113a}$$

   for $x \neq x_i$ and

$$N_A(x_i) = \lim_{x\to x_i} N_A(x), \tag{1.113b}$$

   for all $i \in \{1,\dots,n_{\text{el}}+1\}$.

   The set

$$\{(a,e)\,|\,\mathsf{LG}(a,e) = A\} \tag{1.114}$$

   says that we should seek all pairs (a,e) of shape function index a and element number e that are mapped to basis function index A. In other

words, we should add all shape functions that contribute to basis function N_A. Because each basis function N_A is the sum of some shape functions, the set in 1.114 is never empty. Finally, if the limit in (1.113b) is not defined, the value of $N_A(x_i)$ is left undefined.

To remember that in performing this special sum we add the values everywhere except at the nodes as in (1.113a), and evaluate limits to find their values at the nodes as in (1.113b), we introduce a special name and symbol for it. We call it the **broken sum**, $\mathring{+}: \mathcal{W}_h \times \mathcal{W}_h \to \mathcal{W}_h$, so that for $f_h, g_h \in \mathcal{W}_h$,

$$(f_h \mathring{+} g_h)(x) = f_h(x) + g_h(x), \qquad x \neq x_i, \tag{1.115a}$$

and

$$(f_h \mathring{+} g_h)(x_i) = \lim_{x \to x_i} (f_h \mathring{+} g_h)(x). \tag{1.115b}$$

Consistently with the new symbol, the broken summation sum is $\mathring{\sum}$.

With this notation, we can write

$$N_A = \mathring{\sum_{\{(a,e)|\mathsf{LG}(a,e)=A\}}} N_a^e. \tag{1.116}$$

Notice that we are now regularly referring to different sets of basis functions in the same context: the basis functions for $\mathcal{W}_h$ and the basis functions for the element spaces $\mathcal{P}^e$, or shape functions. To distinguish them, the basis functions for $\mathcal{W}_h$, whose domain is the entire interval, are called **global basis functions**. Conversely, basis functions for the element spaces $\mathcal{P}^e$, defined only over the element domain, are called **local basis functions**.

☞ Basis functions for the finite element space are called *global basis functions*, and basis functions for an element space are called *local basis functions*, or shape functions. Global basis functions are defined in the entire domain of the problem, while local basis functions are defined only over an element.

As a convention and whenever possible, uppercase letters will be used for indices of global degrees of freedom or basis functions, while local degrees of freedom or basis functions will use indices that are lowercase letters.

Examples:

1.68 **A space with discontinuous functions.** The simplest basis to define a space $\mathcal{W}_h$ over the mesh in Fig. 1.14 is the one in which each shape function defines a single global basis function. In this case, $m = k \times n_{\text{el}} = 2 \times 3 = 6$, and a local-to-global map is a 2×3 matrix that can be defined as

$$\mathsf{LG} = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}. \tag{1.117}$$

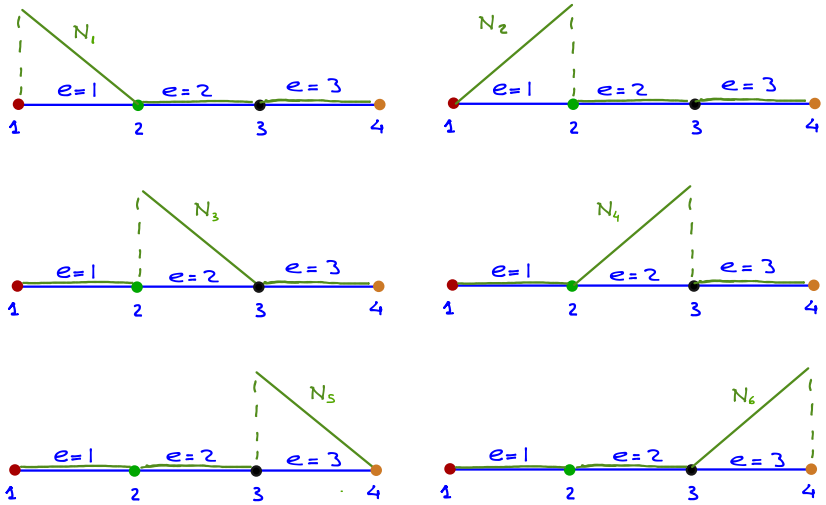
For example,

$$\mathsf{LG}(2,1) = 2,$$
$$\mathsf{LG}(1,3) = 5,$$
$$\mathsf{LG}(2,3) = 6.$$

With this local-to-global map, each global basis function index appears only once, so the basis functions for $\mathcal{W}_h$ can be written as

$$N_1 = N_1^1, \quad N_2 = N_2^1$$
$$N_3 = N_1^2, \quad N_4 = N_2^2$$
$$N_5 = N_1^3, \quad N_6 = N_2^3.$$

A sketch of the global basis functions is



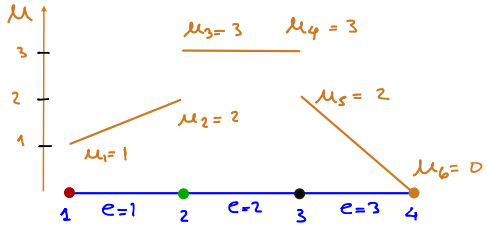
An example of a function defined on this space is

$$u = 1\, N_1 \mathring{+} 2\, N_2 \mathring{+} 3\, N_3 \mathring{+} 3\, N_4 \mathring{+} 2\, N_5 \mathring{+} 0\, N_6,$$

with components in this basis

$$U = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \\ 2 \\ 0 \end{bmatrix}.$$

These components can be interpreted as the one-sided limits of the function at the nodes, as sketched next:



☞ As defined by the broken sum, values at the discontinuities are left undefined, since the limit in (1.115b) does not exist. The two one-sided limits do exist. We will not need to worry about this, since we are going to require functions in the finite element space to be continuous.

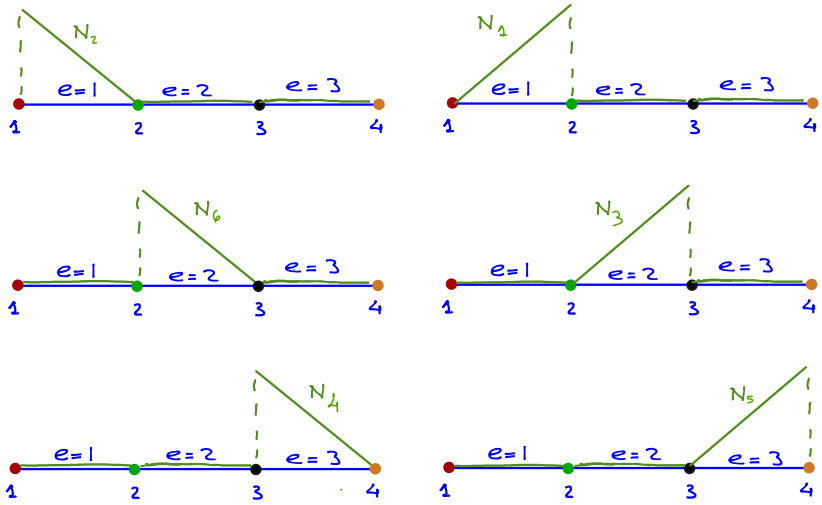
The index we assigned to each basis function of $\mathcal{W}_h$ is immaterial, since $\mathcal{W}_h$ does not change upon altering the name of each basis function. For example, we could have used the following local-to-global map

$$\mathsf{LG} = \begin{bmatrix} 2 & 6 & 4 \\ 1 & 3 & 5 \end{bmatrix}. \tag{1.118}$$

In this case, the global basis functions can be written as

$$\begin{aligned} N_1 = N_2^1, \quad & N_2 = N_1^1 \\ N_3 = N_2^2, \quad & N_4 = N_1^3 \\ N_5 = N_2^3, \quad & N_6 = N_1^2. \end{aligned}$$

A sketch of the global basis functions with this new label is



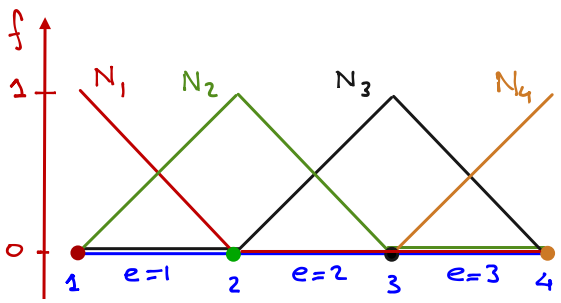
The function u is now written as

$$\begin{aligned} u &= 1\, N_2 \mathring{+} 2\, N_1 \mathring{+} 3\, N_6 \mathring{+} 3\, N_3 \mathring{+} 2\, N_4 \mathring{+} 0\, N_5 \\ &= 2\, N_1 \mathring{+} 1\, N_2 \mathring{+} 3\, N_3 \mathring{+} 2\, N_4 \mathring{+} 0\, N_5 \mathring{+} 3\, N_6, \end{aligned}$$

and its components are

$$U = \begin{bmatrix} 2 \\ 1 \\ 3 \\ 2 \\ 0 \\ 3 \end{bmatrix}.$$

1.69 **The simplest space of continuous functions.** The basis made of hat functions over the mesh in Fig. 1.14 is:



We can then build each basis function of $\mathcal{W}_h$ as a sum of shape functions as follows:

$$\begin{aligned} N_1 &= N_1^1 \\ N_2 &= N_2^1 \mathring{+} N_1^2 \\ N_3 &= N_2^2 \mathring{+} N_1^3 \\ N_4 &= N_2^2. \end{aligned} \tag{1.119}$$

This space has dimension $m = 4$, and the local-to-global map is defined by the following 2×3 matrix:

$$\mathsf{LG} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix}. \tag{1.120}$$

What is then the set

$$\{(a,e) \mid \mathsf{LG}(a,e) = 2\}?$$

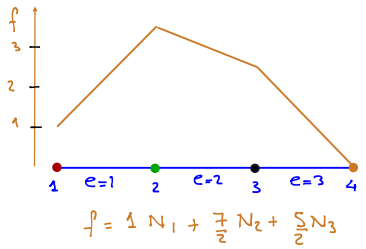
It is $\{(2,1),(1,2)\}$, that is, the rows and columns of the two entries equal to 2 in the LG matrix in (1.120).

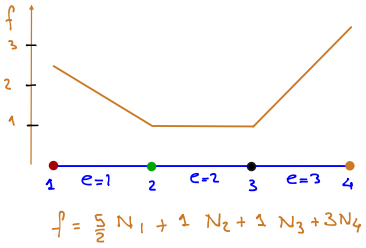
What about the set

$$\{(a,e) \mid \mathsf{LG}(a,e) = 4\}?$$

It is the set $\{(2,3)\}$. You can now check that (1.113a) reduces to (1.119) for this example.

Examples of functions in this space are shown next:





1.70 **A space of continuous piecewise quadratic functions.** Consider the mesh of Fig. 1.14 with P_2-elements, c.f. (1.108). A space $\mathcal{W}_h$ of continuous functions that are polynomials of degree less or equal than 2

over each element can be built with the basis functions sketched in Fig. 1.17. These basis functions can be written as

$$
\begin{aligned}
N_1 &= N_1^1,\\
N_2 &= N_3^1 \mathring{+} N_1^2\\
N_3 &= N_3^2 \mathring{+} N_1^3\\
N_4 &= N_3^3\\
N_5 &= N_2^1,\\
N_6 &= N_2^2,\\
N_7 &= N_2^3.
\end{aligned}
$$

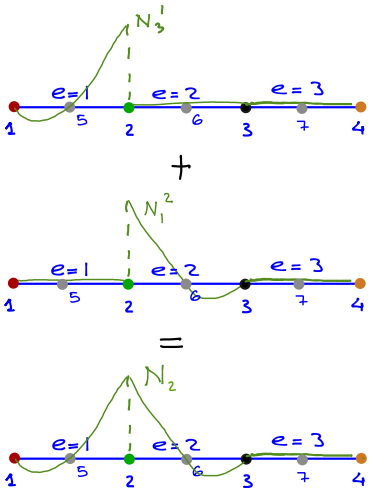


Figure 1.16

For example, the construction of N_2 is sketched in Fig. 1.16.

The space has dimension $m = 7$, and it is obtained from the local-to-global map given by the $k \times n_{\text{el}} = 3 \times 3$ matrix

$$
\mathsf{LG} = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 6 & 7 \\ 2 & 3 & 4 \end{bmatrix}.
$$

For example,

$$
\mathsf{LG}(2,2) = 6, \qquad \mathsf{LG}(3,1) = 2.
$$

It is clear from Examples 1.68 and 1.69 that different local-to-global maps can be defined over the same finite element mesh, in this case a mesh of 3 P_1-elements. Each combination of a finite element mesh and a local-to-global map defines a (potentially different) space $\mathcal{W}_h$. This is a very general and flexible framework, which starting from the definition of finite elements enables the construction of very rich and varied vector spaces of functions.

Similarly to the basis functions, the degrees of freedom of the element space $\mathcal{P}^e$ are labeled **local degrees of freedom**, while those of $\mathcal{W}_h$ are called **global degrees of freedom**. We see next that, by construction, *the local-to-global map is also a map from the index of the local degree of freedom to the index of the global degree of freedom.* Specifically, consider a function $u \in \mathcal{W}_h$ defined by the values $(u_1, \ldots, u_m)$ of the global degrees of freedom, namely,

$$
u = \sum_{A=1}^{m} u_A N_A,
$$

then

$$
u = \mathop{\mathring{\sum}}_{e=1}^{n_{\text{el}}} \mathop{\mathring{\sum}}_{a=1}^{k} u_{\mathsf{LG}(a,e)} N_a^e.
$$

Therefore, for any element e, the function $f^e : \Omega_e \to \mathbb{R}$ defined by

$$
f^e(x) = u(x) \qquad x \in \Omega_e,
$$

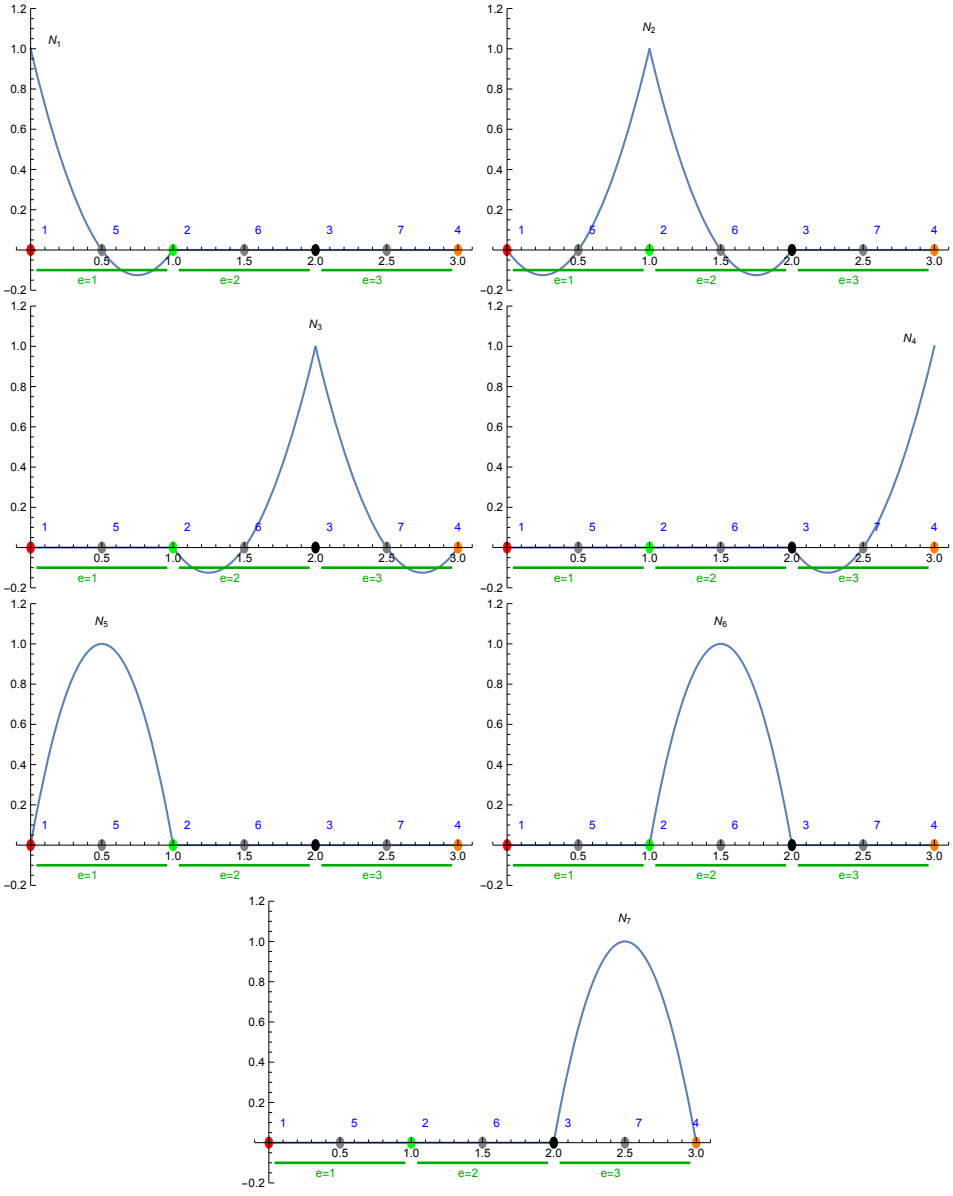


Figure 1.17 Basis functions for a finite element space of 3 P_2-elements.

belongs to $\mathscr{P}^e$. Its components in the local basis are

$$\phi_a^e = u_{\mathsf{LG}(a,e)}. \tag{1.121}$$

The function f^e is called the **restriction** of u to element e.

So each set of values for the global degrees of freedom define a set of values for the local degrees of freedom to describe the same function over an element. The process of obtaining the local degrees of freedom from the global ones through (1.121) is called **localization**.

Map between local and global degrees of freedom (1.121)**.**

Consider $u \in \mathscr{W}_h$, then

$$\begin{aligned}
u(x) &= \sum_{A=1}^{m} u_A N_A(x) \\
&= \sum_{A=1}^{m} u_A \mathop{\sum_{\{(a,e)|\mathsf{LG}(a,e)=A\}}}^{\circ} N_a^e(x) && \text{from (1.116)} \\
&= \sum_{A=1}^{m} \mathop{\sum_{\{(a,e)|\mathsf{LG}(a,e)=A\}}}^{\circ} u_{\mathsf{LG}(a,e)} N_a^e(x) && \text{from definition of } \mathsf{LG} \\
&= \mathop{\sum_{e=1}^{n_{\text{el}}}}^{\circ} \mathop{\sum_{a=1}^{k}}^{\circ} u_{\mathsf{LG}(a,e)} N_a^e(x) && \text{see below}
\end{aligned}$$

The last step uses the fact that in spanning all values of A with the first sum, the two sums together effectively guarantee that all pairs (a,e) will be added exactly once, since $\{1,\ldots,m\}$ is precisely the range of LG, so its pre-image is the entire domain. This is again a consequence of the fact that every shape function contributes to exactly one global basis function, and all global basis functions are built in this way. It is also a consequence of defining global basis functions as sums of shape functions. Had global basis functions been defined as more general linear combinations of shape functions, each local degree of freedom would not be directly equal to a global degree of freedom.

So, because of the construction of the basis functions, it follows that: (a) the function u restricted to element e belongs to $\mathscr{P}^e$, (b) the values of the degrees of freedom of u restricted to element e are $\phi_a^e = u_{\mathsf{LG}(a,e)}$, so the local-to-global map also maps the local degrees of freedom to local ones.

1.4.4 Assembly of the Stiffness Matrix and Load Vector

The computation of the stiffness matrix and load vector generally involves the calculation of integrals over the domain, such as those involved in the bilinear form and linear functional. For example, for model Problem 1.3 with $b(x) = 0$,

$$a(u,v) = \int_\Omega \left[k(x)u'(x)v'(x) + c(x)u(x)v(x)\right]\, dx, \tag{1.122a}$$

$$\ell(v) = k(L)d_L v(L) + \int_\Omega f(x)v(x)\, dx. \tag{1.122b}$$

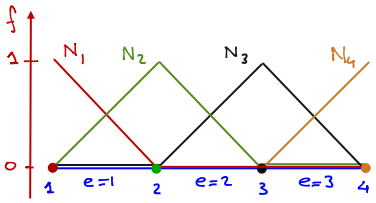


Figure 1.18

This task highlights key functions that elements provide: The decomposition of the domain into elements afford us the ability to decompose integrals over the domain as a sum of integrals over elements, construct *elemental stiffness matrices* and *elemental load vectors*, and *"assemble"* them over the mesh to form the stiffness matrix and load vector of the problem.

Let's have a brief look at the main ideas of what we will be discussing. Consider again the simplest space of continuous functions over a mesh with 3 elements and basis functions in Fig. 1.18. If, for example, we wanted to compute the stiffness matrix entry $K_{33} = a(N_3, N_3)$, we could split the integral in (1.122a) as

$$\begin{aligned} K_{33} &= \int_\Omega \left[k(x)N_3'(x)N_3'(x) + c(x)N_3(x)N_3(x)\right]\,dx \\ &= \underbrace{\int_{\Omega_2} \left[k(x)\left(N_2^2\right)'(x)\left(N_2^2\right)'(x) + c(x)N_2^2(x)N_2^2(x)\right]\,dx}_{K_{22}^2} \\ &\quad + \underbrace{\int_{\Omega_3} \left[k(x)\left(N_1^3\right)'(x)\left(N_1^3\right)'(x) + c(x)N_1^3(x)N_1^3(x)\right]\,dx}_{K_{11}^3} \end{aligned} \tag{1.123}$$

There is no need to compute an integral over element $e = 1$, given that $N_3(x) = 0$ for $x \in \Omega_1$ and hence the value of the integral is zero. Only elements 2 and 3 contribute non-zero values to K_{33}. The contribution of each element, K_{22}^2 and K_{11}^3, are entries of the element stiffness matrices for elements 2 and 3. Over each element we can replace the global basis functions (N_3) by the local ones (N_2^2 and N_1^3), or shape functions. The value of K_{33} is obtained by *accumulating* the contributions to its value by all elements in the mesh. This is what is called *assembling* K_{33}.

For problems with two and three-dimensional domains, computing in this way simplifies the integration problem enormously, since it is only necessary to learn how to compute integrals over each element, rather than over arbitrarily-shaped domains.

Element Stiffness Matrix and Element Load Vector. The element stiffness matrix K^e and element load vector F^e are inspired and emerge from the decomposition of the integrals involved in the definition of the bilinear form and linear functional as sums of integrals over elements. Symbolically, we can write

$$\int_\Omega (\cdot) = \sum_{e=1}^{n_{\text{el}}} \int_{\Omega_e} (\cdot), \tag{1.124}$$

where Ω is the domain over which an integral is performed. Many commonly found bilinear forms and linear functionals can be written as

$$\begin{aligned} a(u,v) &= \sum_{e=1}^{n_{\text{el}}} a^e(u,v) \quad \text{where} \quad a^e(u,v) = \int_{\Omega_e} \ldots\, dx \\ \ell(v) &= \sum_{e=1}^{n_{\text{el}}} \ell^e(v) \quad \text{where} \quad \ell^e(v) = \int_{\Omega_e} \ldots\, dx, \end{aligned} \tag{1.125}$$

For model Problem 1.3 with $b(x) = 0$ and $d_L = 0$ (c.f. (1.122)) this is

$$\begin{aligned} a(u,v) &= \sum_{e=1}^{n_{\text{el}}} \underbrace{\int_{\Omega_e} \left[k(x)u'(x)v'(x) + c(x)u(x)v(x) \right] dx}_{=a^e(u,v)} = \sum_{e=1}^{n_{\text{el}}} a^e(u,v), \\ \ell(v) &= \sum_{e=1}^{n_{\text{el}}} \underbrace{\int_{\Omega_e} f(x)v(x)\, dx}_{\ell^e(v)} = \sum_{e=1}^{n_{\text{el}}} \ell^e(v). \end{aligned} \tag{1.126}$$

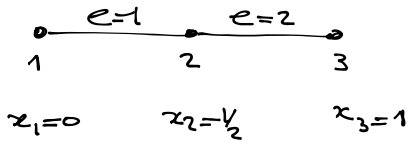
We are now in position to define K^e and F^e as

$K^e_{ab} = a^e(N^e_b, N^e_a)$	Element Stiffness Matrix	(1.127a)
$F^e_a = \ell^e(N^e_a)$	Element Load Vector	(1.127b)

for any $a, b = 1, \ldots, k$. Before we discuss how these are used to construct K and F, let's look at an example.

The definition of the element stiffness matrix and element load vector is also notable for what it does not define: the values of $a^e(N^{e_1}_b, N^{e_2}_a)$ and $\ell^e(N^{e_1}_a)$, for elements e, e_1 and e_2 with $e \neq e_1$ and $e \neq e_2$. The reason for this is that if either $e_1 \neq e$ or $e_2 \neq e$, then both values are identically zero, given that either $N^{e_1}_b$ and/or $N^{e_2}_a$ are zero in Ω_e. Since the values of any basis function N_A in Ω_e are exclusively defined by linear combinations of shape functions in an element, these are the only potentially non-zero contributions to the stiffness matrix or load vector, and hence the only ones included in the element stiffness matrix and element load vector. Therefore, out of the $m \times m$ combinations of basis functions (generally a large number), the only non-zero contributions of $a^e(N_A, N_B)$ are accounted for by the $k \times k$ elemental stiffness matrix (generally a much smaller number).

Example 1.71 We compute the element stiffness matrix K^e and element load vector F^e defined by (1.126) for every element in a simple case. Consider a mesh with two P_1 elements over the interval $[0,1]$, and let $f(x) = 10$, $k(x) = 1$ and $c(x) = 3x$ for $x \in (0,1)$.



We do not yet specify the local-to-global map LG because it is not needed to compute K^e and F^e; it will be needed later to build K and F.

Since we have two elements, we need to compute K^1, K^2, F^1 and F^2. From (1.126) and the values for k, c and f,

$$a^e(u,v) = \int_{\Omega_e} \left[u'(x)v'(x) + 3xu(x)v(x) \right] dx, \qquad \ell^e(v) = \int_e 10v\, dx. \tag{1.128}$$

The shape functions over each element of this finite element mesh are

$$N_1^1(x) = \frac{1/2 - x}{1/2}$$
$$N_2^1(x) = \frac{x}{1/2}$$
$$N_1^2(x) = \frac{1 - x}{1/2}$$
$$N_2^2(x) = \frac{x - 1/2}{1/2}.$$

Notice here the superindex N_a^e with $e = 1, 2$ is the element index, and not exponentiation. To simplify the notation, we will also use $N_{,x}$ to indicate the derivative of N, instead of N'.

The element stiffness matrices follow as:

$$K_{ab}^1 = \int_0^{1/2} N_{a,x}^1 N_{b,x}^1 + 3x N_a^1 N_b^1 \, dx, \qquad K_{ab}^2 = \int_{1/2}^1 N_{a,x}^2 N_{b,x}^2 + 3x N_a^2 N_b^2 \, dx.$$

This results in

$$K^1 = \begin{bmatrix} \int_0^{1/2} \left(-\frac{1}{1/2}\right)\left(-\frac{1}{1/2}\right) + 3x \frac{1/2-x}{1/2}\frac{1/2-x}{1/2}\, dx & \int_0^{1/2} \left(-\frac{1}{1/2}\right)\left(\frac{1}{1/2}\right) + 3x \frac{1/2-x}{1/2}\frac{x}{1/2}\, dx \\ \int_0^{1/2} \left(\frac{1}{1/2}\right)\left(-\frac{1}{1/2}\right) + 3x \frac{x}{1/2}\frac{1/2-x}{1/2}\, dx & \int_0^{1/2} \left(\frac{1}{1/2}\right)\left(\frac{1}{1/2}\right) + 3x \frac{x}{1/2}\frac{x}{1/2}\, dx \end{bmatrix}$$
$$= \begin{bmatrix} \frac{33}{16} & -\frac{31}{16} \\ -\frac{31}{16} & \frac{35}{16} \end{bmatrix},$$
$$K^2 = \begin{bmatrix} \int_{1/2}^1 \left(-\frac{1}{1/2}\right)\left(-\frac{1}{1/2}\right) + 3x \frac{1-x}{1/2}\frac{1-x}{1/2}\, dx & \int_{1/2}^1 \left(-\frac{1}{1/2}\right)\left(\frac{1}{1/2}\right) + 3x \frac{1-x}{1/2}\frac{x-1/2}{1/2}\, dx \\ \int_{1/2}^1 \left(\frac{1}{1/2}\right)\left(-\frac{1}{1/2}\right) + 3x \frac{x-1/2}{1/2}\frac{1-x}{1/2}\, dx & \int_{1/2}^1 \left(\frac{1}{1/2}\right)\left(\frac{1}{1/2}\right) + 3x \frac{x-1/2}{1/2}\frac{x-1/2}{1/2}\, dx \end{bmatrix}$$
$$= \begin{bmatrix} \frac{37}{16} & -\frac{29}{16} \\ -\frac{29}{16} & \frac{39}{16} \end{bmatrix}.$$

The element load vectors are:

$$\begin{aligned} F_a^1 = \int_0^{1/2} 10 N_a^1 \, dx \quad &\Longrightarrow F^1 = \begin{bmatrix} \int_0^{1/2} 10 \frac{1/2-x}{1/2}\, dx \\ \int_0^{1/2} 10 \frac{x}{1/2}\, dx \end{bmatrix} = \begin{bmatrix} 5/2 \\ 5/2 \end{bmatrix}, \\ F_a^2 = \int_{1/2}^1 10 N_a^2 \, dx \quad &\Longrightarrow F^2 = \begin{bmatrix} \int_{1/2}^1 10 \frac{1-x}{1/2}\, dx \\ \int_{1/2}^1 10 \frac{x-1/2}{1/2}\, dx \end{bmatrix} = \begin{bmatrix} 5/2 \\ 5/2 \end{bmatrix}. \end{aligned} \tag{1.129}$$

To encompass the most common types of variational equations, we further need to consider load vectors that have contributions from evaluating test functions on the boundary of the domain. An example of such contribution is found in (1.122b), in the term

$$k(L) d_L v(L).$$

This term cannot be written as an integral over an element, and it involves the value of the test function v at the end of the interval, $x = L$. When terms of this type are present, the linear form of the problem is written as

$$\ell(v) = \sum_{e=1}^{n_{\text{el}}} \ell^e(v) + h_0 v(0) + h_L v(L), \tag{1.130}$$

where h_0 and h_L are two real numbers defined by the bilinear form. For our example in (1.122b), $h_0 = 0$ and $h_L = k(L)d_L$.

For simplicity, we will incorporate the potential contributions of the two boundary terms into the element load vector of the first and last elements of the mesh. Specifically, we modify the definition (1.127a) of the element load vector for $e = 1$ and $e = n_{\text{el}}$ to become

$$\begin{aligned} F_a^1 &= \ell^1(N_a^1) + h_0 N_a^1(0) && \text{Element load vector for } e = 1 \\ F_a^e &= \ell^e(N_a^e) && \text{Element load vector for } e \neq 1, n_{\text{el}} \\ F_a^{n_{\text{el}}} &= \ell^{n_{\text{el}}}(N_a^{n_{\text{el}}}) + h_L N_a^{n_{\text{el}}}(L) && \text{Element load vector for } e = n_{\text{el}}, \end{aligned} \tag{1.131}$$

for $a = 1, \ldots, k$.

Assembly. The construction of the stiffness matrix K and load vector F in terms of the ones from the elements is called the finite element **assembly** operation. It is also called the **Direct Stiffness Method**. It is a result of the way global basis functions are constructed (c.f. (1.113a)). Recall that basis functions in $\mathcal{W}_h$ can be written in terms of the shape functions, or local basis functions, as

$$N_A(x) = \mathop{\mathop{\sum}^{\circ}}_{\{(a,e)\,|\,\mathsf{LG}(a,e)=A\}} N_a^e(x). \tag{1.132}$$

for $A \in \{1, \ldots, m\}$. Based on this relation, we show below that

$$\boxed{F_A = \sum_{e=1}^{n_{\text{el}}} \sum_{\{a\,|\,\mathsf{LG}(a,e)=A\}} F_a^e, \qquad A \in \eta_a} \tag{1.133a}$$

and

$$\boxed{K_{AB} = \sum_{e=1}^{n_{\text{el}}} \sum_{\substack{\{a\,|\,\mathsf{LG}(a,e)=A\} \\ \{b\,|\,\mathsf{LG}(b,e)=B\}}} K_{ab}^e, \qquad A \in \eta_a, B \in \eta} \tag{1.133b}$$

These identities directly connect entries of the stiffness matrix and load vector in active indices' rows to entries in the corresponding contributions from the elements. Specifically, each such entry in K and F is obtained by accumulating the contributions of some elements in the mesh.

Example 1.72 We show next that (1.123) is a result of (1.133b). The local-to-global map for the mesh in Fig. 1.18 is (1.120), namely,

$$\mathsf{LG} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix}.$$

To compute K_{33} from (1.133b), we need to identify the set

$$\{a \,|\, \mathsf{LG}(a, e) = 3\},$$

for each $e \in \{1,2,3\}$, since $A = B = 3$. By inspection of LG, it follows that

$$\{a \mid \mathsf{LG}(a,1) = 3\} = \emptyset,$$
$$\{a \mid \mathsf{LG}(a,2) = 3\} = \{2\},$$
$$\{a \mid \mathsf{LG}(a,3) = 3\} = \{1\}.$$

Replacing with these indices in (1.133b), we obtain

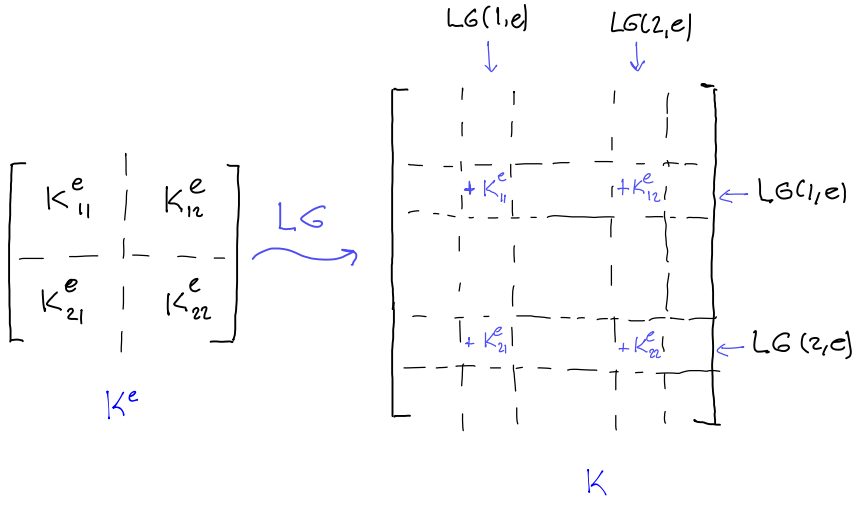
$$K_{33} = K_{22}^2 + K_{11}^3,$$

and we recover (1.123).

This example illustrates that computing a value of K_{AB} for a single pair of indices AB or the value of F_A for a single index A requires searching for those elements that contain indices of local degrees of freedom that are mapped to A and/or B by LG. However, since we want to compute K_{AB} and F_A for *all* active indices A and *all* indices B, it is more efficient to proceed in a different way:

1. Initially set $K = 0$ and $F = 0$.
2. Visit every element e in the mesh to compute K^e and F^e and
3. Add K_{ab}^e to $K_{\mathsf{LG}(a,e)\mathsf{LG}(b,e)}$ for all $a, b\{1,\ldots,k\}$ if $\mathsf{LG}(a,e) \in \eta_a$, and
4. Add F_a^e to $F_{\mathsf{LG}(a,e)}$ for all $a \in \{1,\ldots,k\}$ if $\mathsf{LG}(a,e) \in \eta_a$.

In this way, there is no need to search for which elements contribute to an entry. This is the distinguishing aspect of the assembly. A sketch of the way the local-to-global map defines where to add the element stiffness matrix is shown next:



Example 1.73 Let's revisit Example 1.71 to assemble a stiffness matrix and a load vector.

The mesh contains two elements, and to build the finite element space we need to specify the local-to-global map LG. For a space of continuous functions, this map is

$$\mathsf{LG} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix},$$

and the dimension of $\mathcal{W}_h$ is $m = 3$. Therefore, K is a 3×3 matrix, and F is a 3×1 matrix. Furthermore, we will assume that all indices are active [11], namely, $\eta_a = \eta = \{1, 2, 3\}$.

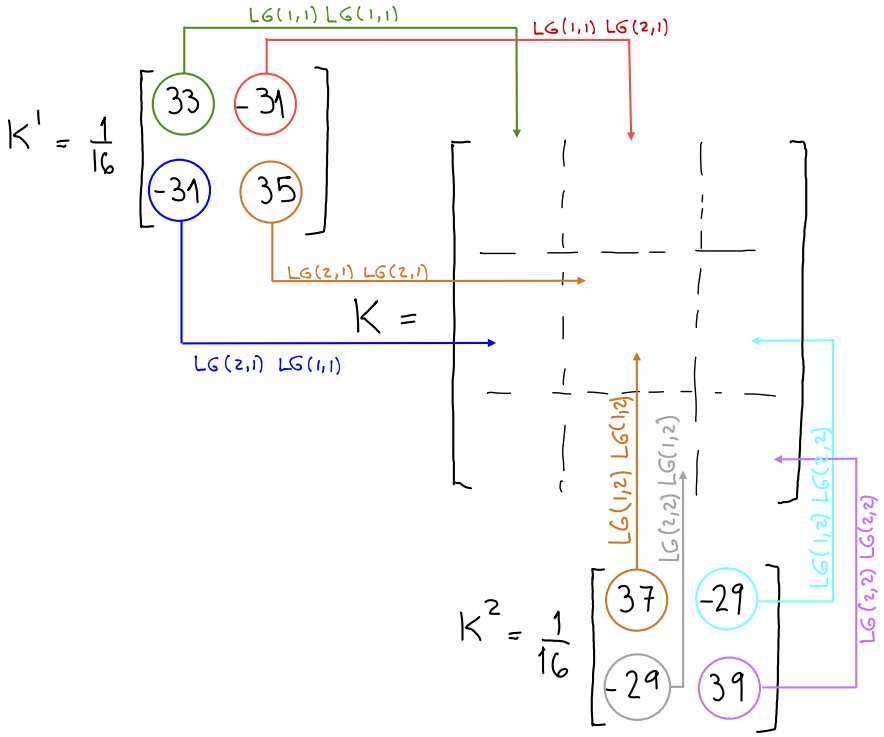
The element stiffness matrices are, from Example 1.71,

$$K^1 = \frac{1}{16}\begin{bmatrix} 33 & -31 \\ -31 & 35 \end{bmatrix} \qquad K^2 = \frac{1}{16}\begin{bmatrix} 37 & -29 \\ -29 & 39 \end{bmatrix},$$

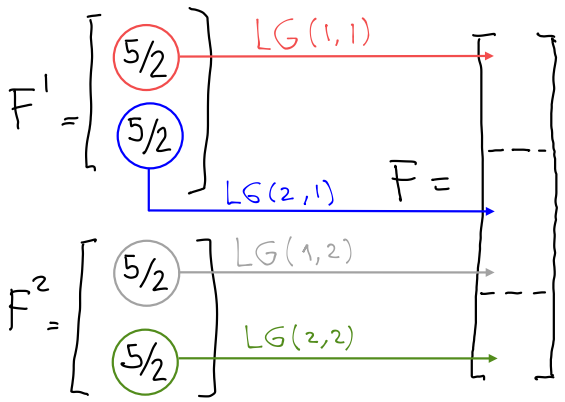
and the element load vectors are

$$F^1 = F^2 = \begin{bmatrix} 5/2 \\ 5/2 \end{bmatrix}.$$

The assembly process is sketched in the following figures



[11] this is the case when the boundary conditions are natural and *homogeneous*, i.e., equal to zero



The results of the assembly are

$$K = \frac{1}{16}\begin{bmatrix} 33 & -31 & 0 \\ -31 & 35+37 & -29 \\ 0 & -29 & 39 \end{bmatrix} = \frac{1}{16}\begin{bmatrix} 33 & -31 & 0 \\ -31 & 72 & -29 \\ 0 & -29 & 39 \end{bmatrix},$$

and

$$F = \begin{bmatrix} 5/2 \\ 5/2+5/2 \\ 5/2 \end{bmatrix} = \begin{bmatrix} 5/2 \\ 5 \\ 5/2 \end{bmatrix}. \tag{1.134}$$

Insofar we have discussed how to assemble the rows of K and F whose indices are active. Rows with constrained indices are still defined by (1.78c), i.e., for $A \in \eta_g$ and $B \in \eta$,

$$K_{AB} = \delta_{AB}, \qquad F_A = \overline{u}_A.$$

Taking this into account, the pseudocode for the assembly procedure is

```
K = 0,  F = 0
FOR e ∈ {1, ..., n_el}
    FOR a ∈ {1, ..., k}
        IF LG(a,e) ∈ η_a
            FOR b ∈ {1, ..., k}
                K(LG(a,e), LG(b,e)) += K^e_ab
            END FOR
            F(LG(a,e)) += F^e_a
        END IF
    END FOR
END FOR
FOR A ∈ η_g
    K(A,A) = 1
    F(A)   = ū_A
END FOR
```

Example 1.74 Let's modify Example 1.73 and assemble K and F by assuming that $\eta_a = \{2,3\}$ and that $\overline{u}_1 = 4$, instead of the original assumption that all indices are active.

The results of the assembly are

$$K = \begin{bmatrix} 1 & 0 & 0 \\ -31/16 & 35/16+37/16 & -29/16 \\ 0 & -29/16 & 39/16 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -31/16 & 72/16 & -29/16 \\ 0 & -29/16 & 39/16 \end{bmatrix},$$

and

$$F = \begin{bmatrix} 4 \\ 5/2+5/2 \\ 5/2 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 5/2 \end{bmatrix}. \tag{1.135}$$

Example 1.75 As a final twist, let's examine the assembly for meshes with a different element type and with an increasing number of elements, so

as to observe the pattern of non-zero entries that emerges. For simplicity, we consider meshes with elements of equal length, and the bilinear form and linear functional in 1.126 with $f(x) = k(x) = c(x) = 1$. In this way, all element stiffness matrices and element load vectors are the same, since none of the three functions depends on x.

Since $\Omega = (0,1)$, the vertices for this mesh are at $x_i = (i-1)h$ for $i = 1,\dots,n_{\text{el}}+1$ and $h = 1/n_{\text{el}}$. We will consider a finite element space made of continuous piecewise quadratic functions (P_2-elements), and as customary, we add a node at the midpoint of each elements to indicate the third local degree of freedom. Thus, we set $x_{n_{\text{el}}+1+i} = (x_i + x_{i+1})/2$ for $i = 1,\dots,n_{\text{el}}$. Figure 1.16 shows the mesh for $n_{\text{el}} = 3$.

The element stiffness matrix is a 3×3 matrix, the load vector has length 3, and are computed as

$$K^e_{ab} = \int_{x_e}^{x_{e+1}} \left[N^e_{a,x}(x) N^e_{b,x}(x) + N_a(x) N_b(x) \right] dx,$$
$$F^e_a = \int_{x_e}^{x_{e+1}} N_a(x)\, dx,$$

where the shape functions are those in (1.108). We explicitly show the computation for two entries of K^e, and leave the rest for the reader to verify. From (1.108), we will use that $x^e_1 - x^e_2 = h$, $x^e_3 - x^e_1 = x^e_2 - x^e_3 = h/2$, and that

$$\begin{aligned} N^e_1\left(x^e_1+\xi h\right) &= (\xi-1)(2\xi-1), & N^e_{1,x}(x^e_1+h\xi) &= \frac{4\xi-3}{h} \\ N^e_3\left(x^e_1+\xi h\right) &= 4(1-\xi)\xi, & N^e_{3,x}(x^e_1+h\xi) &= \frac{4-8\xi}{h} \end{aligned} \tag{1.136}$$

for $\xi \in [0,1]$. Then,

$$\begin{aligned} K^e_{11} &= \int_{x_e}^{x_{e+1}} \left[N^e_{1,x}(x) N^e_{1,x}(x) + N_1(x) N_1(x) \right] dx \\ &= \int_0^1 \left[N^e_{1,x}(x_e+\xi h) N^e_{1,x}(x_e+\xi h) + N_1(x_e+\xi h) N_1(x_e+\xi h) \right] h\, d\xi \qquad \text{Change of variables } \xi = \tfrac{x-x_e}{h} \\ &= \int_0^1 \left[\frac{(4\xi-3)^2}{h^2} + (\xi-1)^2(2\xi-1)^2 \right] h\, d\xi \qquad \text{From (1.136)} \\ &= \frac{7}{3h} + \frac{2h}{15}. \\ K^e_{13} &= \int_{x_e}^{x_{e+1}} \left[N^e_{1,x}(x) N^e_{3,x}(x) + N_1(x) N_3(x) \right] dx \\ &= \int_0^1 \left[N^e_{1,x}(x_e+\xi h) N^e_{3,x}(x_e+\xi h) + N_1(x_e+\xi h) N_3(x_e+\xi h) \right] h\, d\xi \qquad \text{Change of variables } \xi = \tfrac{x-x_e}{h} \\ &= \int_0^1 \left[\frac{(4\xi-3)(4-8\xi)}{h^2} - 4(\xi-1)^2(2\xi-1)\xi \right] h\, d\xi \qquad \text{From (1.136)} \\ &= -\frac{8}{3h} + \frac{h}{15}. \end{aligned}$$

Proceeding with the computation, the element stiffness matrix is then

$$K^e = \begin{bmatrix} \frac{7}{3h} + \frac{2h}{15} & \frac{1}{3h} - \frac{h}{30} & -\frac{8}{3h} + \frac{h}{15} \\ \frac{1}{3h} - \frac{h}{30} & \frac{7}{3h} + \frac{2h}{15} & -\frac{8}{3h} + \frac{h}{15} \\ -\frac{8}{3h} + \frac{h}{15} & -\frac{8}{3h} + \frac{h}{15} & \frac{16}{3h} + \frac{8h}{15} \end{bmatrix}, \tag{1.137}$$

and the element load vector is

$$F^e = \begin{bmatrix} \frac{h}{6} \\ \frac{h}{6} \\ \frac{2h}{3} \end{bmatrix}. \tag{1.138}$$

Next, we assemble the stiffness matrix and load vector for a mesh with 3, 6, and 9 elements of the same length, with $\eta_c = \{1\}$ and $\overline{u}_1 = -1$. As always, we need to decide how to index the global degrees of freedom. In this case, it is customary to adopt the number of the node as the index of the global degree of freedom, so the local-to-global maps for each case are

$$\begin{aligned} n_{\text{el}} = 3, \quad \mathsf{LG} &= \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 5 & 6 & 7 \end{bmatrix} \\ n_{\text{el}} = 6, \quad \mathsf{LG} &= \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 4 & 5 & 6 & 7 \\ 8 & 9 & 10 & 11 & 12 & 13 \end{bmatrix} \\ n_{\text{el}} = 9, \quad \mathsf{LG} &= \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 \end{bmatrix}. \end{aligned} \tag{1.139}$$

When $n_{\text{el}} = 3$, $h = 1/3$, the stiffness matrix is

$$K = \frac{1}{90} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 89 & 1268 & 89 & 0 & -718 & -718 & 0 \\ 0 & 89 & 1268 & 89 & 0 & -718 & -718 \\ 0 & 0 & 89 & 634 & 0 & 0 & -718 \\ -718 & -718 & 0 & 0 & 1456 & 0 & 0 \\ 0 & -718 & -718 & 0 & 0 & 1456 & 0 \\ 0 & 0 & -718 & -718 & 0 & 0 & 1456 \end{bmatrix},$$

and the (transpose of the) force vector is

$$F^{\mathsf{T}} = \frac{1}{18} \begin{bmatrix} -18 & 2 & 2 & 1 & 4 & 4 & 4 \end{bmatrix}.$$

When $n_{\text{el}} = 6$, $h = 1/6$, the stiffness matrix is

$$K = \frac{1}{180} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 359 & 5048 & 359 & 0 & 0 & 0 & 0 & -2878 & -2878 & 0 & 0 & 0 & 0 \\ 0 & 359 & 5048 & 359 & 0 & 0 & 0 & 0 & -2878 & -2878 & 0 & 0 & 0 \\ 0 & 0 & 359 & 5048 & 359 & 0 & 0 & 0 & 0 & -2878 & -2878 & 0 & 0 \\ 0 & 0 & 0 & 359 & 5048 & 359 & 0 & 0 & 0 & 0 & -2878 & -2878 & 0 \\ 0 & 0 & 0 & 0 & 359 & 5048 & 359 & 0 & 0 & 0 & 0 & -2878 & -2878 \\ 0 & 0 & 0 & 0 & 0 & 359 & 2524 & 0 & 0 & 0 & 0 & 0 & -2878 \\ -2878 & -2878 & 0 & 0 & 0 & 0 & 0 & 5776 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2878 & -2878 & 0 & 0 & 0 & 0 & 0 & 5776 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2878 & -2878 & 0 & 0 & 0 & 0 & 0 & 5776 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2878 & -2878 & 0 & 0 & 0 & 0 & 0 & 5776 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2878 & -2878 & 0 & 0 & 0 & 0 & 0 & 5776 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2878 & -2878 & 0 & 0 & 0 & 0 & 0 & 5776 \end{bmatrix},$$

$$K = \frac{1}{270}\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
809 & 11348 & 809 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6478 & -6478 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 809 & 11348 & 809 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6478 & -6478 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 809 & 11348 & 809 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6478 & -6478 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 809 & 11348 & 809 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6478 & -6478 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 809 & 11348 & 809 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6478 & -6478 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 809 & 11348 & 809 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6478 & -6478 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 809 & 11348 & 809 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6478 & -6478 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 809 & 11348 & 809 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6478 & -6478 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 809 & 5674 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6478 \\
-6478 & -6478 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12976 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -6478 & -6478 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12976 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -6478 & -6478 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12976 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -6478 & -6478 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12976 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -6478 & -6478 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12976 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -6478 & -6478 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12976 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -6478 & -6478 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12976 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -6478 & -6478 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12976 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6478 & -6478 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12976
\end{bmatrix},$$

Figure 1.19 Stiffness matrix for $n_{\text{el}} = 9$ in Example 1.75.

and the (transpose of the) force vector is

$$F^{\mathsf{T}} = \frac{1}{36}\begin{bmatrix} -36 & 2 & 2 & 2 & 2 & 2 & 1 & 4 & 4 & 4 & 4 & 4 & 4 \end{bmatrix}.$$

Finally, when $n_{\text{el}} = 9$, $h = 1/9$, the stiffness matrix is shown in Fig. 1.19 and the (transpose of the) force vector is

$$F^{\mathsf{T}} = \frac{1}{54}\begin{bmatrix} -54 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 1 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \end{bmatrix}.$$

The reader may want to follow this example closely to understand how the assembly works.

Derivation of Assembly Formulas (1.133)**.**

In addition to the way in which global basis functions are constructed, in the following, we will use the facts that: (a) shape functions are zero outside the elements over which they are defined, and (b) the integrals over each element in the definition of ℓ^e and a^e will therefore be zero when computed for a shape function of another element. More precisely, for $a \in \{1,\dots,m\}$ and $e, e', e'' \in \{1,\dots,n_{\text{el}}\}$,

$$N_a^e(x) = 0 \text{ if } x \notin \Omega_e, \text{ from where, } \begin{cases} \ell^e\left(N_a^{e'}\right) = 0 & \text{if } e \neq e', \\ a^e(N_a^{e''}, N_b^{e'}) = 0 & \text{if } e' \neq e \text{ or } e'' \neq e. \end{cases} \tag{1.140}$$

We examine how we arrive to (1.133a) first. For simplicity, we proceed by assuming

that $h_0 = 0$, since it reduces the bookkeeping needed in the derivation. For $A \in \eta_a$,

$$
\begin{aligned}
F_A &= \ell(N_A) && \text{from (1.78}c\text{)}\\
&= \sum_{e=1}^{n_{\text{el}}} \ell^e(N_A) + h_L N_A(L) && \text{from (1.130)}\\
&= \sum_{e=1}^{n_{\text{el}}} \ell^e\left(\mathring{\sum_{\{(a,e')|\mathsf{LG}(a,e')=A\}}} N_a^{e'}\right) + h_L \mathring{\sum_{\{(a,e)|\mathsf{LG}(a,e)=A\}}} N_a^e(L) && \text{from (1.132)}\\
&= \sum_{e=1}^{n_{\text{el}}} \sum_{\{(a,e')|\mathsf{LG}(a,e')=A\}} \ell^e\left(N_a^{e'}\right) + \mathring{\sum_{\{(a,e)|\mathsf{LG}(a,e)=A\}}} h_L N_a^e(L) && \text{linearity}\\
&= \sum_{e=1}^{n_{\text{el}}} \sum_{\{a|\mathsf{LG}(a,e)=A\}} \ell^e(N_a^e) + \sum_{\{a|\mathsf{LG}(a,n_{\text{el}})=A\}} h_L N_a^{n_{\text{el}}}(L), && \text{from (1.140)}\\
&= \sum_{e=1}^{n_{\text{el}}-1} \sum_{\{a|\mathsf{LG}(a,e)=A\}} \ell^e(N_a^e) + \sum_{\{a|\mathsf{LG}(a,n_{\text{el}})=A\}} \left[\ell^{n_{\text{el}}}(N_a^{n_{\text{el}}}) + h_L N_a^{n_{\text{el}}}(L)\right] &&\\
&= \sum_{e=1}^{n_{\text{el}}} \sum_{\{a|\mathsf{LG}(a,e)=A\}} F_a^e. && \text{from (1.131).}
\end{aligned}
$$

It should be evident from this derivation that the equality in the last line does not change if $h_0 \neq 0$. This proves (1.133a).

Similarly, to obtain (1.133b), for $A \in \eta_a, B \in \eta$,

$$
\begin{aligned}
K_{AB} &= a(N_B, N_A) && \text{from (1.78}c\text{)}\\
&= \sum_{e=1}^{n_{\text{el}}} a^e(N_B, N_A) && \text{from (1.125)}\\
&= \sum_{e=1}^{n_{\text{el}}} a^e\left(\mathring{\sum_{\{(b,e'')|\mathsf{LG}(b,e'')=B\}}} N_b^{e''}, \mathring{\sum_{\{(a,e')|\mathsf{LG}(a,e')=A\}}} N_a^{e'}\right) && \text{from (1.132)}\\
&= \sum_{e=1}^{n_{\text{el}}} \sum_{\{(a,e')|\mathsf{LG}(a,e')=A\}} \sum_{\{(b,e'')|\mathsf{LG}(b,e'')=B\}} a^e(N_b^{e''}, N_a^{e'}) && \text{bilinearity}\\
&= \sum_{e=1}^{n_{\text{el}}} \sum_{\substack{\{a|\mathsf{LG}(a,e)=A\}\\ \{b|\mathsf{LG}(b,e)=B\}}} a^e(N_b^e, N_a^e) && \text{from (1.140)}\\
&= \sum_{e=1}^{n_{\text{el}}} \sum_{\substack{\{a|\mathsf{LG}(a,e)=A\}\\ \{b|\mathsf{LG}(b,e)=B\}}} K_{ab}^e && \text{from (1.127}a\text{).}
\end{aligned}
$$

1.4.4.1 Symmetrization of the Stiffness Matrix

When the bilinear from is symmetric, it is possible to manipulate the stiffness matrix to obtain a symmetric matrix. Symmetric matrices are needed when some iterative solvers for the linear system are adopted, such as Conjugate Gradients. Additionally, symmetric matrices can be stored with less memory, or more efficiently. Problems in structural mechanics, such as a elasticity, have a symmetric bilinear form and can benefit from a symmetric stiffness matrix.

Given a stiffness matrix K such that (a) $K_{AB} = K_{BA}$ for $A \in \eta_a, B \in \eta_a$ and, (b) $K_{AB} = \delta_{AB}$ if $A \notin \eta_a$, we can construct a symmetric matrix K^S and load vector F^S such that U is the solution of both

$$KU = F \quad \text{and} \quad K^S U = F^S. \tag{1.141}$$

It is then possible to solve $K^S U = F^S$ to find U, instead of $KU = F$.

Any stiffness matrix that emerges from a variational method of the form in Problem 1.2 satisfies condition (a) if the bilinear form is symmetric, and satisfies condition (b) automatically, c.f. (1.78c).

The symmetric stiffness matrix and associated load vector follow as

$$K^S_{AB} = \begin{cases} K_{AB} & \text{if } A \in \eta_a, B \in \eta_a, \\ \delta_{AB} & \text{otherwise.} \end{cases} \tag{1.142a}$$

$$F^S_A = \begin{cases} F_A - \sum_{B \in \eta_g} K_{AB} F_B & \text{if } A \in \eta_a, \\ F_A & \text{otherwise} \end{cases} \tag{1.142b}$$

Example 1.76 Let's symmetrize the matrix in Example 1.74. Notice that the matrix in the earlier example, Example 1.73, is already symmetric. The stiffness matrix and load vector from Example 1.74 are

$$K = \begin{bmatrix} 1 & 0 & 0 \\ -31/16 & 72/16 & -29/16 \\ 0 & -29/16 & 39/16 \end{bmatrix}, \qquad F = \begin{bmatrix} 4 \\ 5 \\ 5/2 \end{bmatrix},$$

with $\eta_a = \{2,3\}$.

Before we proceed with the direct construction of the matrix, we look at why it works. The linear system defined by the stiffness matrix and load vector is

$$\begin{array}{llll} 1 \cdot u_1 & +0 \cdot u_2 & +0 \cdot u_3 & = 4, \\ -31/16 \cdot u_1 & +72/16 \cdot u_2 & -29/16 \cdot u_3 & = 5, \\ 0 \cdot u_1 & -29/16 \cdot u_2 & +39/16 \cdot u_3 & = 5/2. \end{array}$$

Because the first line is a constrained index, it defines directly the value of $u_1 = 4$. We can then replace this in the first column of lines 2 and 3, and move them to the right hand side. The linear system can then be written as

$$\begin{array}{llll} 1 \cdot u_1 & +0 \cdot u_2 & +0 \cdot u_3 & = 4, \\ 0 \cdot u_1 & +72/16 \cdot u_2 & -29/16 \cdot u_3 & = 5 + 31/16 \cdot 4, \\ 0 \cdot u_1 & -29/16 \cdot u_2 & +39/16 \cdot u_3 & = 5/2. \end{array}$$

The matrix associated to this linear system is then symmetric. This is what (1.142) is doing.

Notice that the conditions for symmetrization are satisfied. Condition (a) is satisfied because the submatrix formed by the entries (2,2), (2,3), (3,2), and (3,3) is symmetric. Condition (b) is satisfied because the first row is identically zero, except for the diagonal, in which it is equal to 1. This is of course expected, since this matrix emerged from a variational numerical method and the symmetric bilinear form in Example 1.71.

The symmetrized matrix is

$$K^S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 72/16 & -29/16 \\ 0 & -29/16 & 39/16 \end{bmatrix},$$

The associated load vector F^S follows as

$$F^S = \begin{bmatrix} 4 \\ 5 \\ 5/2 \end{bmatrix} - 4 \begin{bmatrix} 0 \\ -31/16 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ 51/4 \\ 5/2 \end{bmatrix}.$$

To conclude, you can check that both systems lead to the same solution

$$U = \begin{bmatrix} 4 \\ 9116/1967 \\ 8796/1967 \end{bmatrix}.$$

Derivation of Symmetrization Formulas (1.142)

To see that $KU = F$ if and only if $K^S U = F^S$, we consider first the equations for $A \notin \eta_a$. In this case,

$$F_A = K_{AB} u_B = \delta_{AB} u_B = K^S_{AB} u_B = F^S_A. \tag{1.143}$$

Next, we consider the equations for $A \in \eta_a$. We will need to use that from 1.143,

$$F_B = u_B \text{ for } B \notin \eta_a, \tag{1.144}$$

and from (1.142a),

$$K^S_{AB} = 0 \text{ if } A \in \eta_a, B \notin \eta_a. \tag{1.145}$$

Then,

$$\begin{aligned}
0 &= \sum_{B \in \eta} K_{AB} u_B - F_A \\
&= \sum_{B \in \eta_a} K_{AB} u_B + \sum_{B \notin \eta_a} K_{AB} F_B - F_A && \text{from (1.142}a\text{)} \\
&= \sum_{B \in \eta_a} K^S_{AB} u_B - \left(F_A - \sum_{B \in \eta_g} K_{AB} F_B \right) && \text{from (1.144)} \\
&= \sum_{B \in \eta_a} K^S_{AB} u_B - F^S_A && \text{from (1.142}b\text{)} \\
&= \sum_{B \in \eta_a} K^S_{AB} u_B + \sum_{B \notin \eta_a} K^S_{AB} u_B - F^S_A && \text{from (1.145)} \\
&= \sum_{B \in \eta} K^S_{AB} u_B - F^S_A.
\end{aligned}$$

Together with (1.143), this proves that if U satisfies one set of equations, it satisfies the others.

1.4.5 Finite Element Bases and Sparse Stiffness Matrices

A glance at the stiffness matrix computed with P_1-elements in (1.101) reveals that the matrix has non-zero entries only along three of its diagonals, so it is called a **tri-diagonal** matrix. Similarly, an examination of the stiffness matrices in Example 1.75 shows that the only non-zero entries lie along 7 different diagonal directions. This means that as the number of elements grows, the majority of the entries in the matrix are equal to zero. In fact, the number of non-zero entries grows linearly with the number of rows $n_{\text{el}}+1$, since it is proportional to the length of the diagonals. In contrast, the number of zeros grows quadratically with the number of rows, since it corresponds to the rest of the $(n_{\text{el}}+1)^2$ entries. Matrices in which the fraction of non-zero entries is small are called **sparse**.

The sparsity of the matrix has consequences in the computational efficiency of a method. If the matrix is going to be stored in memory[12], it is convenient to store the non-zero entries only. This drastically reduces the memory requirements as the number of elements grows, since the amount of memory needed is $\mathcal{O}(n_{\text{el}})$ instead of $\mathcal{O}({n_{\text{el}}}^2)$ as $n_{\text{el}} \to \infty$. The same scaling law applies to matrix-vector products, which is the main operation used in the solution of linear systems of equations through iterative methods .

For the forthcoming discussion, it is convenient to introduce the definition of support of a function. Given a real-valued function f with domain Ω, the set

$$\text{supp}(f) = \overline{\{x \in \Omega \mid f(x) \neq 0\}} \tag{1.146}$$

is called the **support** of f. In (1.146), the line over set indicates its closure: it says that the support of f is formed by all the points in the set, but also by those that may not be in the set but that can be reached as limits of sequences of points in the set. For the forthcoming discussion, it is enough to think about the support of f as essentially the set of all points in Ω at which f is not equal to zero.

Back to the main discussion then, when are stiffness matrices in the finite element method sparse? For active indices, entries in the stiffness matrix of a variational method are computed as $K_{AB} = a(N_B, N_A)$. The most commonly found scenario is illustrated by the bilinear form in Example 1.55

$$a(u,v) = \int_0^1 u_{,x}(x)\, v_{,x}(x)\, dx. \tag{1.147}$$

In this case, $a(N_B, N_A) = 0$ if $N_{B,x}$ and $N_{A,x}$ are different than zero over non-intersecting regions of the domain, or more precisely, when the intersection of their supports has zero length. For a finite element space of continuous functions

[12] The so-called matrix-free methods never build the matrix, see e.g. CITE

over P_1-elements, $N_{B,x}N_{A,x}$ is a non-zero function only when $|A-B| \leq 1$, that is, when A and B are indices of the same node or of neighboring nodes. This is the reason for the appearance of the tri-diagonal matrix in (1.101), since for each row A, only the entries $K_{A(A-1)}$, K_{AA} and $K_{A(A+1)}$ are non-zero. There are always only at most three non-zero entries per row regardless of the number of elements n_{el}, so the matrix becomes increasingly sparse as n_{el} grows.

A similar scenario is found with the bilinear form of Example 1.75,

$$a(u,v) = \int_0^1 u_{,x}v_{,x} + uv\,dx. \tag{1.148}$$

In this case, if A is the index of a node between elements in the mesh, row A has at most 5 non-zero entries: $K_{AB} \neq 0$ only if $|A-B| \leq 1$ (indices of the neighboring nodes or the same node), $B = A+n_{\text{el}}$, or $B = A+n_{\text{el}}+1$ (indices of the neighboring nodes in the middle of an element); see Fig. 1.19. Alternatively, if A is the index of a node in the middle of an element, then $K_{AB} \neq 0$ only if $B = A-n_{\text{el}}$ (the index of the node on its right) or $B = A-n_{\text{el}}-1$ (the index node of the node on its left). As in the previous example, the number of non-zero entries per row is the same for all values of n_{el}, so the matrix becomes increasingly sparse as n_{el} grows.

These two examples should be contrasted with one of a variational method with non-finite element bases. For example, it is possible to select $\mathcal{W}_h = \mathbb{P}_{m-1}(\Omega)$ for $m \geq 1$, with basis functions $N_A(x) = x^{A-1}$ for $A = 1,\ldots,m$. In this case, $\text{supp}(N_A) = \Omega$ for all A, so the support of each basis function is the entire domain. For any of the two bilinear forms (1.147) or (1.148), choosing this basis leads to non-sparse matrices, as we can expect from the non-empty intersection of the support of any pair of basis functions. Specifically, assuming all indices are active, we have

$$K_{AB} = \int_0^1 (A-1)(B-1)x^{A+B-4}\,dx = \frac{(A-1)(B-1)}{A+B-3}, \qquad \text{for (1.147)}$$

$$K_{AB} = \int_0^1 (A-1)(B-1)x^{A+B-4} + x^{A+B-2}\,dx = \frac{(A-1)(B-1)}{A+B-3} + \frac{1}{A+B-1}. \qquad \text{for (1.148)}$$

These are all non-zero entries (except for the $A = 1$ row or $B = 1$ column for (1.147)). Matrices in which most of the entries are non-zero are called **dense**.

Sufficient Conditions for Stiffness Matrices in the Finite Element Method to be Sparse

The following are sufficient conditions on a finite element basis and the bilinear form to generate a sparse stiffness matrix.

If

(a) $\text{supp}(f) \cap \text{supp}(g) = \emptyset \Longrightarrow a(f,g) = 0$, and

(b) There exists $n_{\text{width}} \in \mathbb{N}$ such that for all $A = 1,\ldots,m$ the number of elements of the set $s_A = \{e \mid \mathsf{LG}(a,e) = A\}$ is less or equal than n_{width}, i.e., $\#s_A \leq n_{\text{width}}$,

then the number of non-zero entries in each row of K is less or equal than $n_{\text{width}} \times k$.

To see this, notice first that $\text{supp}(N_A) = \cup_{e \in s_A} \Omega_e$, since these are all of the elements in which a shape function is added to form N_A. Therefore, due to the first condition, $K_{AB} = 0$ if $s_A \cap s_B = \emptyset$. For each element $e \in s_A$, there are at most k other basis functions whose support includes Ω_e, given that each shape function is added to one and only one global basis function, c.f., 1.4.3. Thus, due to the second condition, the set $\{B \in \{1, \ldots, m\} \mid s_A \cap s_B \neq \emptyset\}$ has at most $n_{\text{width}} \times k$ elements, and hence each row of K has at most that number of non-zero elements.

Let's briefly discuss the conditions and the implications of this result. The first condition is satisfied by essentially all commonly found bilinear forms. The second condition examines the local-to-global map to request global basis functions to be formed by adding at most n_{width} shape functions, regardless of the number of elements in the mesh. Because the number of non-zero entries is less or equal than $n_{\text{width}} \times k$, a quantity that does not change as n_{el} grows, the associated stiffness matrix grows increasingly sparse as $n_{\text{el}} \to \infty$.

We conclude this section with a couple of final remarks:

- The use of methods that generate non-sparse matrices can be convenient when they lead to numerical solutions of similar accuracy with fewer global degrees of freedom than a finite element method, such as in some boundary element methods or spectral methods.

- The property of the finite element basis functions to be non-zero only in a few elements is typically referred to in the literature by stating that *finite element basis functions have compact support*. It is worth reflecting a bit on the meaning of this statement. Mathematically, a function f is said to have **compact support** if $\text{supp}(f)$ is a bounded set. Hence, any function whose support is included in the domain $\Omega = (0, L)$ of our example has compact support. It is clear then that a literal interpretation of this statement does not capture that the support of a finite element basis function is included in the union of at most a fixed number of elements, regardless of the number of elements in the mesh.

1.5 Elliptic Fourth-Order Problems

The goal of this section is to show how to build variational methods for fourth-order problems. For this purpose, we will reenact the same sequence of the previous sections. We will first formulate a class of general fourth-order problems and obtain a variational equation. Then we will deduce and implement a variational method and build an adequate finite element space.

A key insight that we will gain from this example will be the convenience of having additional smoothness of the finite element functions, so that a variational method for the problem is trivially consistent.

1.5.1 The Differential Equation

The elliptic fourth-order problems that most frequently appear in applications are of the general form

$$\big(q(x)\,u''(x)\big)'' - \big(k(x)\,u'(x)\big)' + c(x)\,u(x) = f(x)\,. \tag{1.149}$$

The field under study is, as before, u, and q, k, c and f are the coefficients. This class of equations is broad enough to model very interesting problems.

Examples:

1.77 *Euler-Bernoulli beam equation:* Consider a rectilinear beam with Young modulus $E(x)$ whose cross-section has moment of inertia relative to a neutral horizontal axis $I(x)$. Let $u(x)$ model the (small) vertical deflection of the beam when it is subjected to a vertical load (per unit length) $f(x)$. Then the vertical displacement must satisfy the equation

$$\big(E(x)\,I(x)\,u''(x)\big)'' = f(x) \tag{1.150}$$

at all points x for the beam to be in static equilibrium. In this case $k(x) = c(x) = 0$ and $q(x) = E(x)\,I(x)$ is the *bending rigidity* of the beam.

1.78 *Diffuse-interfaces in material science:* When a material separates into two distinct phases, the process is often modeled by a *diffuse interphase* equation formulated in terms of a *concentration variable* u. A classical example is the *Cahn-Hilliard equation.* Let us consider here a steady-state, linearized form of this equation, which reads

$$q\,u'''' - k\,u'' = f\,, \tag{1.151}$$

where q represents a *diffusion coefficient* and the other coefficients arise from the linearization. This is certainly a particular case of (1.149).

1.79 *Image denoising:* In this application an input image u_0 (a function of x) is to be transformed so as to remove its noise. To do this, the "denoised image" $u(x)$ is defined as the solution to

$$\big(q(x)\,u''\big)'' + u = u_0 \tag{1.152}$$

for a carefully chosen coefficient $q(x)$ (in fact, $q(x)$ is often a function of u itself, but this would turn the equation *nonlinear*, which is outside the scope of the chapter).

For simplicity, in the following we restrict our attention to the case $k = 0$.

It is known from the theory of ordinary differential equations that, if $q(x) \neq 0$, four boundary conditions are required to completely specify u. Of the many possibilities, the class of problems we consider (*elliptic* problems) impose *two conditions on each boundary point of the domain* $\Omega = (0, L)$. The most popular boundary conditions are:

- **Clamped conditions**, which specify the values of $u(0)$ *and* of $u'(0)$ (and/or of $u(L)$ *and* of $u'(L)$, depending on which boundary is considered).
- **Applied load conditions**, which specify the values of $u''(0)$ and of $u'''(0)$ (and/or $u''(L)$ and $u'''(L)$).

To study both types of conditions simultaneously, let us consider a problem with clamped conditions at $x = 0$ and applied load conditions at $x = L$. Other combinations are easy to understand by analogy. The problem is:

Problem 1.5. (Fourth-order problem) *Given the coefficients q, c and f (as functions of x), together with the boundary constants g_0, d_0, m_L and n_L, find a continuously differentiable function $u : \Omega \to \mathbb{R}$ satisfying*

$$\big(q(x)u''(x)\big)'' + c(x)u(x) = f(x) \qquad \forall\, x \in \Omega \tag{1.153a}$$
$$u(0) = g_0 \tag{1.153b}$$
$$u'(0) = d_0 \tag{1.153c}$$
$$u''(L) = m_L \tag{1.153d}$$
$$u'''(L) = n_L \tag{1.153e}$$

Examples:

1.80 *The simplest beam problem* is a homogeneous cantilever beam (q independent of x, $c = 0$) with no distributed load $f = 0$, clamped horizontally at $x = 0$ (i.e., $u(0) = u'(0) = 0$) and with a vertical force W and a bending moment T applied at $x = L$. The strong problem simplifies to

$$u''''(x) = 0,\ \forall x \in \Omega, \quad u(0) = u'(0) = 0, \quad u''(L) = \frac{T}{q}, \quad u'''(L) = -\frac{W}{q}.$$

Since $u'''' = 0$, the solution is necessarily a cubic polynomial, and because of the clamped conditions at $x = 0$ it must be of the form

$$u(x) = c_1\, x^2 + c_2\, x^3\ .$$

It only remains to calculate c_1 and c_2 so that the boundary conditions at $x = L$ are satisfied. The exact solution is

$$u(x) = \frac{T + WL}{2q}\, x^2 - \frac{W}{6q}\, x^3\ .$$

The tip displacement is, in particular,

$$u(L) = \frac{T}{2q} L^2 + \frac{W}{3q} L^3\ .$$

1.81 Other interesting boundary conditions are **elastic support** conditions, which are an analog to the Robin conditions discussed in Section 1.1.1. Their expression is (considering the boundary at $x = 0$)

$$u''(0) - \alpha_0\, u'(0) = \beta_0\,, \qquad u'''(0) + \gamma_0\, u(0) = \delta_0\,, \tag{1.154}$$

where α_0, β_0, γ_0 and δ_0 are given constants. Similarly to the second-order case, making α_0 very large *de facto* imposes the value of $u'(0)$, and making γ_0 very large imposes $u(0)$. In the limit $\alpha_0 \to +\infty$, $\gamma_0 \to +\infty$ we end up with *clamped* conditions. On the other hand, when $\alpha_0 = \gamma_0 = 0$ we recover the *applied load* conditions.

It is always important to check that the problem we are considering is well posed, in the sense that one can expect to have a unique solution which depends continuously on the coefficients and boundary conditions. In this case we have:

Theorem 1.3. *Under the hypotheses that the coefficients q and c are piecewise smooth and non-negative, with $q(x) \geq q_{\min} > 0$, $\forall x$, and that $\int_0^L |f(x)|\,dx < \infty$, Problem 1.5 is well posed.*

1.5.2 A Variational Equation

Let us follow the procedure outlined in §1.1.2.3 to determine a suitable variational equation for Problem 1.5. The residual is

$$r(x) = \big(q(x)\,u''(x)\big)'' + c(x)\,u(x) - f(x). \tag{1.155}$$

If u is the solution, then $r(x) = 0$ for all x, and thus for any smooth function $v(x)$ it must hold that

$$0 = \int_0^L r(x)\,v(x)\,dx = \int_0^L \Big[\big(q(x)\,u''(x)\big)'' + c(x)\,u(x) - f(x)\Big]\; v(x)\,dx. \tag{1.156}$$

After distributing the product inside the bracket, we integrate by parts twice the term $\int_0^L (q u'')''\, v\,dx$ so that the differentiation order is balanced between u and v, arriving at

$$\begin{aligned}\int_0^L \big[q(x)\,u''(x)\,v''(x) + c(x)\,u(x)\,v(x)\big]\;dx = \int_0^L f(x)\,v(x)\,dx \\ - (q u''' + q' u'')(L)\,v(L) + (q u''' + q' u'')\,v(0) \\ + q(L)\,u''(L)\,v'(L) - q(0)\,u''(0)\,v'(0).\end{aligned} \tag{1.157}$$

The second and third lines of (1.157) contain the boundary values of u and v and their derivatives. We next replace with the boundary conditions we have

information about and that appear in the boundary terms. In this case, these are the values of $u''(L) = m_L$ and $u'''(L) = n_L$, to get

$$\int_0^L \left[q(x)u''(x)v''(x) + c(x)u(x)v(x)\right] dx = \int_0^L f(x)v(x)\,dx - (q(L)n_L + q'(L)m_L)v(L) + (qu''' + q'u'')v(0) + q(L)m_L v'(L) - q(0)u''(0)v'(0). \quad (1.158)$$

Since we do not know the values of $u'''(0)$ and $u''(0)$, we will request $v(0) = 0$ and $v'(0) = 0$ in the definition of the test space.

The **natural boundary conditions** for this problem are then $u''(L) = m_L$ and $u'''(L) = n_L$, while $u(0) = g_0$ and $u'(0) = d_0$ need to be considered **essential boundary conditions**.

A variational equation that the solution u of Problem 1.5 satisfies is

$$a(u, v) = \ell(v) \qquad \forall v \in \mathcal{V}, \quad (1.159a)$$

where

$$a(u, v) = \int_0^L \left[q(x)u''(x)v''(x) + c(x)u(x)v(x)\right] dx, \quad (1.159b)$$

$$\ell(v) = \int_0^L f(x)v(x)\,dx - \left(q(L)n_L + q'(L)m_L\right) v(L) + q(L)m_L v'(L) \quad (1.159c)$$

and

$$\mathcal{V} = \left\{v\colon [0, L] \to \mathbb{R} \text{ smooth} \mid v(0) = 0 \text{ and } v'(0) = 0\right\}. \quad (1.159d)$$

Notice that **the bilinear form $a(\cdot,\cdot)$ is symmetric**. You can check that the Euler-Lagrange equations of this variational equation are (1.153a), (1.153d) and (1.153e).

For completeness, the weak form of Problem 1.5 reads:

Problem 1.6. *(Weak form of Problem 1.5) Let the trial space be*

$$\mathcal{S} = \left\{v\colon [0, L] \to \mathbb{R} \textit{ smooth } \mid v(0) = g_0 \textit{ and } v'(0) = d_0\right\}$$

Find $u \in \mathcal{S}$ such that $a(u, v) = \ell(v)$ for all $v \in \mathcal{V}$.

In this weak form, the test space $\mathcal{V}$ is the direction of the trial space $\mathcal{S}$, so it has the general structure of Problem 1.4, with $\mathcal{W} = \{v\colon [0, L] \to \mathbb{R} \text{ smooth}\}$. To see this, let $v, w \in \mathcal{S}$ and $z = v - w$. Then, $z(0) = v(0) - w(0) = g_0 - g_0 = 0$ and $z'(0) = v'(0) - w'(0) = d_0 - d_0 = 0$, and thus $z \in \mathcal{V}$.

In the context of Euler-Bernouilli beams, it so happens that $-\left(q(L)n_L + q'(L)m_L\right)$ equals the applied force load W at L, while $q(L)m_L$ equals the applied torque T at L. So, an equivalent form of writing (1.159c) is

$$\ell(v) = \int_0^L f(x)v(x)\,dx + W\,v(L) + T\,v'(L)\,. \quad (1.160)$$

1.5.3 A Variational Method

A variational method for variational equation (1.159a) is constructed as in Problem 1.2. To this end, we need to select a vector space $\mathcal{W}_h$, and set

$$\mathcal{S}_h = \{w_h \in \mathcal{W}_h \mid w_h(0) = g_0 \text{ and } w_h'(0) = d_0\}$$
$$\mathcal{V}_h = \{w_h \in \mathcal{W}_h \mid w_h(0) = 0 \text{ and } w_h'(0) = 0\}.$$

The variational method then reads:

$$\textit{Find } u_h \in \mathcal{S}_h \textit{ such that } a(u_h, v_h) = \ell(v_h) \textit{ for all } v_h \in \mathcal{V}_h.$$

The bilinear and linear forms are those in (1.159b) and (1.159c)). Let's see how we follow the solution procedure in §1.3.2 to obtain a solvable linear system of equations that allows us to compute u_h.

1.5.3.1 A Variational Method with Global Polynomials

To proceed, we need a finite-dimensional space $\mathcal{W}_h$. For this example, we set $\mathcal{W}_h = \mathbb{P}_k([0, L])$ (polynomials of degree $\leq k$). The next step is to choose a basis of $\mathcal{W}_h$ of which a subset is a basis of $\mathcal{V}_h$. Let

$$N_1(x) = 1, \quad N_2(x) = x, \quad \ldots \quad N_{k+1}(x) = x^k\,,$$

then

$$\{N_1, N_2, \ldots, N_{k+1}\} \qquad \text{is a basis of } \mathcal{W}_h$$

and

$$\{N_3, N_4, \ldots, N_{k+1}\} \qquad \text{is a basis of } \mathcal{V}_h.$$

In fact, N_1 and N_2 are the only two basis functions of $\mathcal{W}_h$ that do not satisfy $N_a(0) = N_a'(0) = 0$. Therefore, the set of active indices is $\eta_a = \{3, \ldots, k+1\}$ and the set of constrained indices is $\eta_g = \{1, 2\}$.

By direct inspection, we see that

$$m = \dim \mathcal{W}_h = k+1 \qquad \text{and} \qquad n = \dim \mathcal{V}_h = m-2 = k-1\,.$$

Following the same reasoning as in §1.3.2, we write

$$u_h(x) = u_1 N_1(x) + u_2 N_2(x) + \ldots + u_m N_m(x) \tag{1.161}$$

so that u_b ($b = 1, \ldots, m$) are the components of the numerical solutions and the algebraic unknowns of our problem. We can then choose $\overline{u}_h \in \mathcal{S}_h$ to be $\overline{u}_h = g_0 + d_0 x$, or $\overline{u}_h = g_0 + d_0 x + x^3$, for example. In both cases, $u_1 = g_0$ and $u_2 = d_0$.

The linear system thus reads

$$\begin{aligned} u_1 &= g_0 \\ u_2 &= d_0 \\ \sum_{b=1}^{m} u_a\, a(N_b, N_3) &= \ell(N_3) \\ \sum_{b=1}^{m} u_a\, a(N_b, N_4) &= \ell(N_4) \\ \cdots &\quad \cdots \\ \sum_{b=1}^{m} u_a\, a(N_b, N_m) &= \ell(N_m) \end{aligned}$$

Written in matrix form, we have

$$K\,U = F\,,$$

with

$$K_{ab} = \begin{cases} \delta_{ab} & \text{if } a = 1 \text{ or } a = 2 \\ a(N_b, N_a) & \text{if } a > 2 \end{cases},$$

and

$$F_a = \begin{cases} g_0 & \text{if } a = 1 \\ d_0 & \text{if } a = 2 \\ \ell(N_a) & \text{if } a > 2. \end{cases}.$$

Example 1.82 To compute actual numbers, let us consider the "simplest beam problem" introduced in Example 1.80, so that $q(x) = q$ is constant, $c = f = 0$, $m_L = T/q$ and $n_L = -W/q$. Let us choose $k = 4$, so that $W_h = \mathbb{P}_4([0, L])$ and thus $m = 5$.

The elements K_{ab} can be calculated by straightforward integration. For $a > 2$ and $b > 2$ this results in

$$\begin{aligned} a(N_b, N_a) &= \int_0^L q\, N_b'' N_a''\, dx \\ &= \int_0^L q(b-1)(b-2)x^{b-3}(a-1)(a-2)x^{a-3}\, dx \\ &= \frac{q(b-1)(b-2)(a-1)(a-2)L^{a+b-5}}{a+b-5} \end{aligned}$$

and $a(N_b, N_a) = \delta_{ab}$ otherwise. Thus,

$$K = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 4qL & 6qL^2 & 8qL^3 \\ 0 & 0 & 6qL^2 & 12qL^3 & 18qL^4 \\ 0 & 0 & 8qL^3 & 18qL^4 & (144/5)qL^5 \end{pmatrix}.$$

The value of the determinant of K is $\det K = 9.6\, q^3\, L^9$.

The entries of the load vector F for $a > 2$ are

$$F_a = W\, N_a(L) + T\, N_a'(L) = W\, L^{a-1} + T\,(a-1)\, L^{a-2} \qquad (a > 2),$$

giving

$$F = \begin{pmatrix} 0 \\ 0 \\ WL^2 + 2TL \\ WL^3 + 3TL^2 \\ WL^4 + 4TL^3 \end{pmatrix}.$$

It only remains to solve $KU = F$. The solution reads

$$U = \begin{pmatrix} 0 \\ 0 \\ (T + WL)/2q \\ -W/(6q) \\ 0 \end{pmatrix}$$

(as can be easily checked by substitution) implying that the numerical solution is

$$u_h = 0 \cdot 1 + 0 \cdot x + \frac{T + WL}{2q} \cdot x^2 - \frac{W}{6q} \cdot x^3 + 0 \cdot x^4\,,$$

coincident with the exact solution computed in Example 1.80. This is always true for the a variational method that is consistent: **if the exact solution lies in $\mathscr{S}_h$, then the solution of the variational method coincides with the exact solution**.

1.5.3.2 The Convenience of a Smoother Space

Some variational methods for fourth-order problems require the use of a space $\mathscr{W}_h$ made of C^1 functions, that is, continuous functions with continuous derivatives. In particular, the variational method based on variational equation (1.159a) does. The reason for this is *consistency*, and if $\mathscr{W}_h$ is not made of C^1 functions, such method is not consistent. As a result, the method may not converge to the exact solution of Problem 1.5, a fact that we will see in §3.

To evaluate consistency, we need to see if the exact solution of Problem 1.5 is also a solution of the method, or Problem 1.2. Specifically, is

$$a(u, v_h) = \ell(v_h) \qquad \forall\, v_h \in \mathscr{V}_h?$$

To see that this is not necessarily true, we consider the case with $T = W = 0$, $q(x) = 1$ and $c(x) = 0$, for simplicity. This implies that $m_L = n_L = 0$, and hence that $u''(L) = u'''(L) = 0$. In this case, we need to check if for all $v_h \in \mathscr{V}_h$

$$\int_0^L u''(x)\, v_h''(x)\, dx = \int_0^L f(x)\, v_h(x)\, dx. \tag{1.162}$$

If $\mathcal{V}_h$ contains discontinuous functions, or functions with discontinuous derivatives, then the integration by parts formula in §1.45 needs to be applied. For simplicity, let's assume that functions in $\mathcal{V}_h$ can have a discontinuity in the function or its derivative at $x = L/2$ only. In this case, integrating by parts the left hand side of (1.162) to transfer a derivative to u we obtain

$$\begin{aligned}
&\int_0^L f(x)\, v_h(x)\, dx \\
&\qquad = \underbrace{u''(L)}_{=0} v_h'(L) - u''(0)\, \underbrace{v_h'(0)}_{=0} + [\![u''(x)\, v_h'(x)]\!]_{x=L/2} - \int_0^L u'''(x)\, v_h'(x)\, dx \qquad (v_h \in \mathcal{V}_h \text{ and } m_L = 0) \\
&\qquad = u''(L/2)\, [\![v_h'(x)]\!]_{x=L/2} - \int_0^L u'''(x)\, v_h'(x)\, dx. \qquad (u'' \text{ is continuous at } x = L/2)
\end{aligned}$$

Integrating by parts once more we get

$$\begin{aligned}
\int_0^L f(x)\, v_h(x)\, dx &= u''(L/2)\, [\![v_h'(x)]\!]_{x=L/2} - \underbrace{u'''(L)}_{=0} v_h(L) + u'''(0)\, \underbrace{v_h(0)}_{=0} \qquad (v_h \in \mathcal{V}_h \text{ and } n_L = 0) \\
&\quad - [\![u'''(x)\, v_h(x)]\!]_{x=L/2} + \int_0^L u''''(x)\, v_h(x)\, dx \\
\int_0^L \underbrace{(f(x) + u''''(x))}_{=0}\, v_h(x)\, dx &= u''(L/2)\, [\![v_h'(x)]\!]_{x=L/2} - u'''(L/2)\, [\![v_h(x)]\!]_{x=L/2}. \qquad (u''' \text{ is continuous at } x = L/2 \text{ and (1.153a)})
\end{aligned}$$

Therefore, the method would be consistent if and only if

$$0 = u''(L/2)\, [\![v_h'(x)]\!]_{x=L/2} - u'''(L/2)\, [\![v_h(x)]\!]_{x=L/2}$$

for *all* $v_h \in \mathcal{V}_h$. Since we can select the jumps of v_h or v_h' arbitrarily by selecting appropriate functions in $\mathcal{V}_h$, we conclude that the method is consistent if and only if $u''(L/2) = 0$ and $u'''(L/2)$. Since this is not necessarily true for the exact solution of Problem 1.5, this variational method is not necessarily consistent.

Instead, if $\mathcal{W}_h$ is made of C^1 functions, $[\![v_h'(x)]\!] = [\![v_h(x)]\!] = 0$ for all $x \in (0, L)$, and hence this method would be consistent.

It is possible to formulate consistent methods with a space $\mathcal{W}_h$ that contains functions that are not necessarily C^1, but it requires a different variational equation, inspired for example by the interior penalty method in Example 1.16.

1.5.4 The Simplest C^1 Finite Element Space

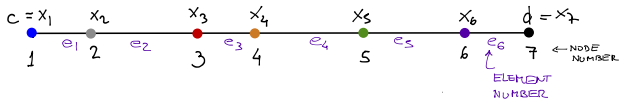
Let us now introduce the simplest **finite element space** that can approximate the fourth-order Problem 1.5, known as the **Hermite piecewise cubic space**. We will denote it here as H_3-space. We follow the same steps as in §1.4.1, but notice that some important differences arise.

Steps:

1. **Build the mesh of the domain.** Let the domain of the problem be the interval $[0, L]$. We partition the domain into $n_{\text{el}} \in \mathbb{N}$ intervals by selecting vertices $\{x_i\}_{i=1,\ldots,n_{\text{el}}+1}$ such that

$$0 = x_1 < \ldots < x_{n_{\text{el}}+1} = L. \tag{1.163}$$

   Interval $[x_i, x_{i+1}]$ is the element domain for element i, for $i = 1, \ldots, n_{\text{el}}$.



2. **Build basis functions.** Remarkably, the "hat functions" $\{N_a\}$ introduced earlier are not useful to approximate fourth-order problems with a variational method based on variational equation (1.159a), since hat functions have discontinuous derivatives, and therefore the method would not be consistent.

   More generally, finite element spaces built with Lagrange P_k-elements will contain functions whose first derivative is discontinuous across element boundaries, and hence cannot be used to built $\mathcal{W}_h$. We thus proceed to introduce a *new* set of basis functions $\{H_k(x), k = 1, 2, \ldots, 2n_{\text{el}} + 2\}$, which are known as **Hermite basis functions**. Their most important feature is that their first derivative is continuous in $[0, L]$, which is why we say that they generate a C^1 **finite element space**. The number of Hermite basis functions equals *twice* the number n_{vert} of vertices. Their second derivative is discontinuous along element boundaries but this does not preclude us from computing the necessary integrals, in the same way that we were able to compute $\int_0^L N_a'(x) N_b'(x)\, dx$ for the hat functions although their first derivatives are discontinuous.

   The Hermite basis functions are **piecewise cubic polynomials**, but not any piecewise cubic polynomial since only a subset of them (in fact, a subspace) is contained in $C^1([0, L])$. The dimension of the vector space Z of piecewise cubic polynomials in a mesh of n_{el} elements is $4\, n_{\text{el}}$, because there are 4 linearly independent cubic polynomials *per element*. The subspace that consists only of C^1 functions, $\mathcal{W}_h = Z \cap C^1([0, L])$, incorporates 2 linear restrictions (continuity of function and derivative) at each of the $n_{\text{el}} - 1$ inter-element boundaries and thus has dimension $m = 4n_{\text{el}} - 2(n_{\text{el}} - 1) = 2n_{\text{el}} + 2 = 2n_{\text{vert}}$. We thus need *two basis functions per vertex* to provide a basis for W_h.

   The Hermite basis functions, for $a = 1, \ldots, m = 2n_{\text{vert}}$, are defined as

a) Odd-numbered basis functions:

$$H_{2a-1}(x) = \begin{cases} 0 & \text{if } x < x_{a-1} \\ -2\left(\frac{x-x_{a-1}}{x_a-x_{a-1}}\right)^3 + 3\left(\frac{x-x_{a-1}}{x_a-x_{a-1}}\right)^2 & \text{if } x_{a-1} \le x < x_a \\ 1 & \text{if } x = x_a \\ 2\left(\frac{x-x_a}{x_{a+1}-x_a}\right)^3 - 3\left(\frac{x-x_a}{x_{a+1}-x_a}\right)^2 + 1 & \text{if } x_a < x \le x_{a+1} \\ 0 & \text{if } x_{a+1} < x \end{cases} \tag{1.164}$$

b) Even-numbered basis functions:

$$H_{2a}(x) = \tag{1.165}$$
$$= \begin{cases} 0 & \text{if } x < x_{a-1} \\ \left[\left(\frac{x-x_{a-1}}{x_a-x_{a-1}}\right)^3 - \left(\frac{x-x_{a-1}}{x_a-x_{a-1}}\right)^2\right](x_a - x_{a-1}) & \text{if } x_{a-1} \le x < x_a \\ 0 & \text{if } x = x_a \\ \left[\left(\frac{x-x_a}{x_{a+1}-x_a}\right)^3 - 2\left(\frac{x-x_a}{x_{a+1}-x_a}\right)^2 + \left(\frac{x-x_a}{x_{a+1}-x_a}\right)\right](x_{a+1} - x_a) & \text{if } x_a < x \le x_{a+1} \\ 0 & \text{if } x_{a+1} < x \end{cases}$$

As before, when $a = 1$, $x \in [0, L]$ implies that we only have the case $x \ge x_a = x_1 = 0$ and when $a = n_{\text{vert}}$, $x \in [0, L]$ implies that we only have the case $x \le x_a = x_{n_{\text{vert}}} = L$. These functions are plotted below for a mesh of three elements and four vertices, $x_0 = 0$, $x_1 = 1$, $x_2 = 2$.

By direct inspection of (1.164)-(1.165) it is evident that these functions are piecewise cubic polynomials. It is not difficult to check that they belong to $C^1([0, L])$ for any valid mesh positions ($x_{a+1} - x_a$ must be strictly positive for all $a = 1, \ldots, n_{\text{vert}} - 1$). You can also verify the following properties:

(a) They add up to 1, i.e., $\sum_{a=1}^{2n_{\text{vert}}} H_a(x) = 1$ for $x \in [0, L]$.

(b) They are linearly independent.

(c) They have **compact support**. The support of H_{2a-1} and H_{2a} is the interval $[x_{a-1}, x_{a+1}]$.

(d) The odd-numbered functions satisfy $H_{2a-1}(x_a) = 1$ and $H'_{2a-1}(x_a) = 0$ at their associated vertices, while $H_{2a-1}(x_b) = H'_{2a-1}(x_b) = 0$ for all other vertices $x_b \neq x_a$.

(e) The even-numbered functions satisfy $H_{2a}(x_b) = 0$ for all vertices x_b of the mesh. On the other hand, their derivative is one at the associated vertex and zero at all other vertices, i.e., $H'_{2a}(x_a) = 1$ and $H'_{2a}(x_b) = 0$ for all $x_b \neq x_a$.

The H_3-space is the vector space of all linear combinations of the functions $H_1, H_2, \ldots, H_{2n_{\text{vert}}}$. From the previous properties, we conclude that the H_3-space is exactly the same space as $\mathcal{W}_h = Z \cap C^1([0, L])$ (the piecewise cubic polynomials that are C^1), i.e.,

$$\mathcal{W}_h = \text{span}\left(H_1, H_2, \ldots, H_{2n_{\text{vert}}}\right), \tag{1.166}$$

and that $H_1, H_2, \ldots, H_{2n_{\text{vert}}}$ is a basis of $\mathcal{W}_h$.

Further, from items (d) and (e) above, we know that for w_h arbitrary in $\mathcal{W}_h$,

$$w_h(x) = c_1 H_1(x) + c_2 H_2(x) + \ldots + c_{2n_{\text{vert}}} H_{2n_{\text{vert}}}(x) \,,$$

it holds that

$$\begin{cases} c_1 &= w_h(x_1), \\ c_2 &= w'_h(x_1), \\ c_3 &= w_h(x_2), \\ c_4 &= w'_h(x_2), \\ \ldots & \ldots \\ c_{2n_{\text{vert}}-1} &= w_h(x_{n_{\text{vert}}}), \\ c_{2n_{\text{vert}}} &= w'_h(x_{n_{\text{vert}}}). \end{cases} \tag{1.167}$$

The arbitrary coefficients c_i ($i = 1, \ldots, m = 2n_{\text{vert}}$) are the **degrees of freedom of the space**. The *odd* degree of freedom c_{2k-1} is the value of w_h at vertex x_k. The *even* degree of freedom c_{2k} is the value of w'_h at vertex x_k. Because of these, this finite element mesh has a node at each vertex. Henceforth, we will refer to them as nodes.

3. **Build $\mathcal{V}_h$.** The boundary conditions at $x = L$, which are *natural* boundary conditions, have already been incorporated into the weak form (1.159a). On the other hand, the boundary conditions at $x = 0$ are *essential* and thus need to be imposed explicitly in the definition of the trial and test spaces, $\mathcal{S}_h$ and $\mathcal{V}_h$.

   How do we do that? We define both $\mathcal{S}_h$ and $\mathcal{V}_h$ as **suitable subsets** of $\mathcal{W}_h$. Let w_h be an arbitrary function in $\mathcal{W}_h$,

$$w_h(x) = c_1 H_1(x) + c_2 H_2(x) + \ldots + c_{2n_{\text{vert}}} H_{2n_{\text{vert}}}(x) \,. \tag{1.168}$$

   Now, since $x_1 = 0$, we have that

$$w_h(0) = c_1, \qquad \text{and} \qquad w'_h(0) = c_2 \,. \tag{1.169}$$

Because the value of $w_h(0)$ involves only c_1 and that of $w_h'(0)$ involves only c_2 it is straightforward to build the trial and test spaces for our problem. The basis was purposefully designed to make things easy.

Functions v_h belonging to the **test space** $\mathcal{V}_h$, to begin with, need to satisfy $v_h(0) = v_h'(0) = 0$. These two linear restrictions are automatically satisfied if

$$\mathcal{V}_h = \left\{ v_h \in \mathcal{W}_h | v_h(0) = v_h'(0) = 0 \right\} = \text{span}\left(H_3, H_4, \ldots, H_{2n_{\text{vert}}} \right), \tag{1.170}$$

meaning that $\mathcal{V}_h$ consists of all functions of the form

$$v_h(x) = c_3 H_3(x) + c_4 H_4(x) + \ldots + c_{2n_{\text{vert}}} H_{2n_{\text{vert}}}(x) \,, \tag{1.171}$$

or, equivalently, all functions in $\mathcal{W}_h$ that have $c_1 = c_2 = 0$.

Functions z_h in the **trial space** $\mathcal{S}_h$, in turn, need to satisfy $z_h(0) = g_0$ and $z_h'(0) = d_0$. From (1.169) we know that this takes place if and only if $c_1 = g_0$ and $c_2 = d_0$. This means that the functions in $\mathcal{S}_h$ can be written as

$$z_h(x) = g_0 H_1(x) + d_0 H_2(x) + c_3 H_3(x) + \ldots + c_{2n_{\text{vert}}} H_{2n_{\text{vert}}}(x) \,, \tag{1.172}$$

where the coefficients $c_3, \ldots, c_{2n_{\text{vert}}}$ are arbitrary. Another way of writing the definition of $\mathcal{S}_h$ is

$$\begin{aligned} \mathcal{S}_h &= \left\{ z_h \in \mathcal{W}_h | z_h = g_0 H_1 + d_0 H_2 + v_h,\, v_h \in \mathcal{V}_h \right\} \\ &= g_0 H_1 + d_0 H_2 + \mathcal{V}_h \,. \end{aligned} \tag{1.173}$$

If we go back to the discussion that led to (1.72b), we see that (1.173) *invites* us to select $\overline{u}_h = g_0 H_1 + d_0 H_2$ (we could add any linear combination of $H_3, \ldots, H_{2n_{\text{vert}}}$ to it, but … adding nothing is simpler and better!)

The set of indices of all basis functions in $\mathcal{W}_h$ is

$$\eta = \{1, 2, \ldots, 2n_{\text{vert}}\},$$

the set of constrained indices is

$$\eta_g = \{1, 2\},$$

and the set of active indices follows as

$$\eta_a = \{3, 4, \ldots, 2n_{\text{vert}}\}.$$

The number of indices in $\eta \setminus \eta_g$ is n, the dimension of $\mathcal{V}_h$, and is thus the number of linearly independent equations generated by our weak form when $v_h \in \mathcal{V}_h$. In our example, this number is $2n_{\text{vert}} - 2$ (i.e., $m - 2$). Adding the two equations coming from the boundary conditions $u_h(0) = c_1 = g_0$ and $u_h'(0) = c_2 = d_0$ we arrive at m equations with m unknowns.

4. **Compute K and F.** Let the finite element solution be denoted by

$$u_h(x) = u_1 H_1(x) + u_2 H_2(x) + \ldots + u_m H_m(x) \,, \tag{1.174}$$

and let U be the column vector of its coefficients,

$$U = (u_1, u_2, \ldots, u_m)^T \,.$$

Inserting (1.174) into $a(u_h, v_h)$, particularizing for $v_h = H_a$, with $a \in \eta_a$ and incorporating the essential boundary conditions, we have that, for $a, b \in \eta = \{1, \ldots, m\}$,

$$K_{ab} = \begin{cases} \delta_{ab} & \text{if } a \in \eta_g = \{1,2\}, \\ a(H_b, H_a) & \text{if } a \in \eta_a \,. \end{cases} \tag{1.175}$$

For the load vector,

$$F_a = \begin{cases} g_0 & \text{if } a = 1 \,, \\ d_0 & \text{if } a = 2 \,, \\ \ell(H_a) & \text{if } a \in \eta_a \,. \end{cases} \tag{1.176}$$

Example 1.83 Fourth-order problem with uniform mesh and constant coefficients. Let us carry out the explicit computations corresponding to a uniform mesh with mesh size $h = L/n_{\text{el}}$. For this, we bring back the definitions of $a(\cdot,\cdot)$ and $\ell(\cdot)$ from Problem 1.6 and those of the basis functions from (1.164)-(1.165). We assume that $q > 0$, $c \geq 0$ and f are constants.

A direct calculation shows that, for $a = 1, \ldots, n_{\text{vert}}$,

$$H''_{2a-1}(x) = \begin{cases} 0 & \text{if } x < x_{a-1} \\ -12(x - x_{a-1})/h^3 + 6/h^2 & \text{if } x_{a-1} < x < x_a \\ 12(x - x_a)/h^3 - 6/h^2 & \text{if } x_a < x < x_{a+1} \\ 0 & \text{if } x > x_{a+1} \end{cases}$$

$$H''_{2a} = \begin{cases} 0 & \text{if } x < x_{a-1} \\ 6(x - x_{a-1})/h^2 - 4/h & \text{if } x_{a-1} < x < x_a \\ 6(x - x_a)/h^2 - 2/h & \text{if } x_a < x < x_{a+1} \\ 0 & \text{if } x > x_{a+1} \end{cases}$$

The next step is the (tedious) computation of all the system matrix

and load vector components. The result is the following:

$$K = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ C_1 & C_2 & C_3 & 0 & C_1 & -C_2 & 0 & 0 & \dots & 0 & 0 \\ -C_2 & C_4 & 0 & C_5 & C_2 & C_4 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & C_1 & C_2 & C_3 & 0 & C_1 & -C_2 & \dots & 0 & 0 \\ 0 & 0 & -C_2 & C_4 & 0 & C_5 & C_2 & C_4 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & C_1 & C_2 & C_3 & 0 & C_1 & -C_2 \\ 0 & 0 & \dots & 0 & 0 & -C_2 & C_4 & 0 & C_5 & C_2 & C_4 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & C_1 & C_2 & C_3/2 & C_6 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & -C_2 & C_4 & C_6 & C_5/2 \end{pmatrix}$$

where

$$\begin{aligned} C_1 &= -\frac{12q}{h^3} + \frac{9ch}{70} \\ C_2 &= -\frac{6q}{h^2} + \frac{13ch^2}{420} \\ C_3 &= \frac{24q}{h^3} + \frac{26ch}{35} \\ C_4 &= \frac{2q}{h} - \frac{ch^3}{140} \\ C_5 &= \frac{8q}{h} + \frac{2ch^3}{105} \\ C_6 &= -\frac{6q}{h^2} - \frac{11ch^2}{210} \end{aligned}$$

and

$$F = \begin{pmatrix} g_0 \\ d_0 \\ fh \\ 0 \\ fh \\ 0 \\ \dots \\ fh \\ 0 \\ fh/2 - qn_L \\ qm_L \end{pmatrix}.$$

5. **Solve and Compute the Finite Element Solution.** We now solve the system $KU = F$, and then build the finite element solution as $u_h(x) = \sum_{a\in\eta} u_a H_a(x)$.

   Taking $L = 1$, $q = 10$, $c = 0$, $f = -1$, $g_0 = 0$, $d_0 = 0$, $n_L = 0$ and $m_L = 0$ we get the conditions of a bar of constant cross section, clamped on the left

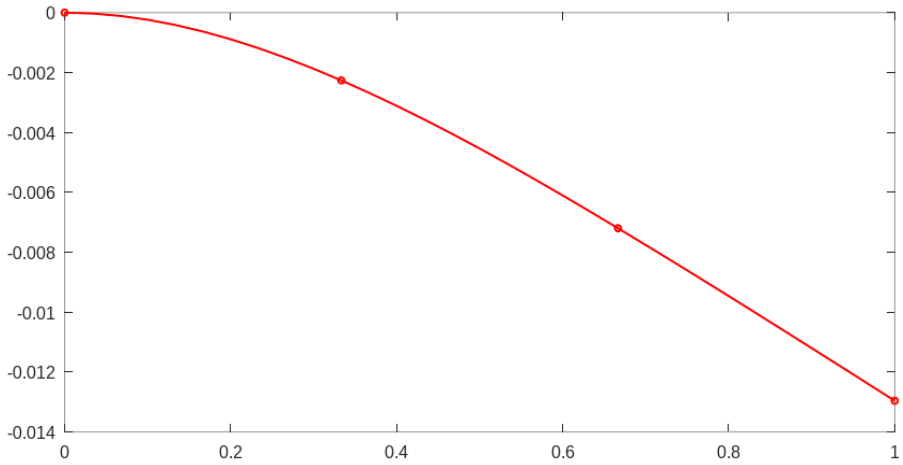


Figure 1.20 The finite element solution u_h corresponding to Equation 1.177.

boundary and free on the right boundary, with uniform load. For a small mesh with just three elements ($h = 1/3$, $n_{\text{vert}} = 4$, and thus $m = 8$) we get the solution

$$U = \begin{bmatrix} 0 \\ 0 \\ -2.2634e-03 \\ -1.2037e-02 \\ -7.2016e-03 \\ -1.6667e-02 \\ -1.2963e-02 \\ -1.7593e-02 \end{bmatrix}$$

and hence

$$u_h(x) = -0.0023H_3(x) - 0.012H_4(x) - 0.0072H_5(x) \\ - 0.0167H_6(x) - 0.013H_7(x) - 0.0176H_8(x). \quad (1.177)$$

This function is plotted in Fig. 1.20.

1.5.5 The Cubic Hermite Finite Element

The basis functions $H_1,\ldots,H_{2n_{\text{vert}}}$ introduced in (1.164)-(1.165) can also be viewed as generated by the following finite element:

$$\Omega_e = \left[x_1^e, x_2^e\right], \tag{1.178}$$

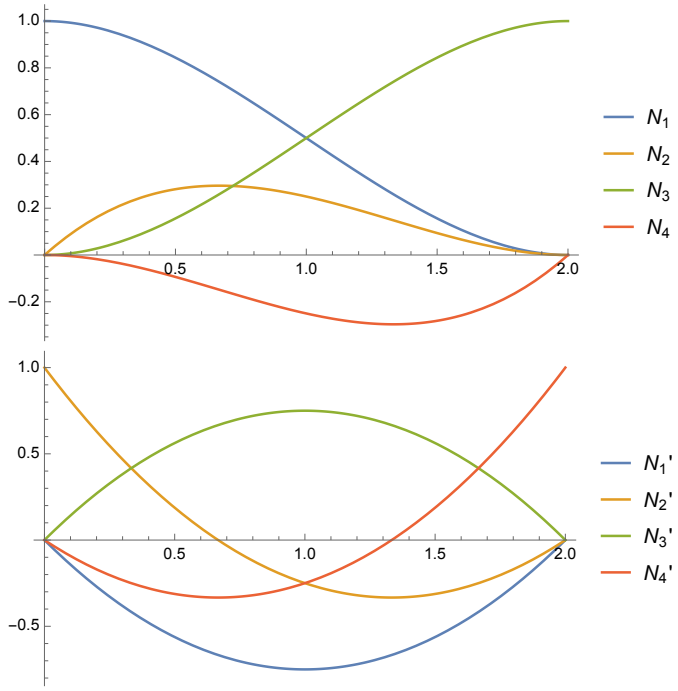
$$N_1^e(x) = \left(\frac{x_2^e - x}{x_2^e - x_1^e}\right)^2 \left(1 + 2\frac{x - x_1^e}{x_2^e - x_1^e}\right), \tag{1.179}$$

$$N_3^e(x) = \left(\frac{x_1^e - x}{x_1^e - x_2^e}\right)^2 \left(1 + 2\frac{x - x_2^e}{x_1^e - x_2^e}\right), \tag{1.180}$$

$$N_2^e(x) = \left(\frac{x_2^e - x}{x_2^e - x_1^e}\right)^2 \left(x - x_1^e\right), \tag{1.181}$$

$$N_4^e(x) = \left(\frac{x_1^e - x}{x_1^e - x_2^e}\right)^2 \left(x - x_2^e\right). \tag{1.182}$$

These functions and their derivatives are plotted next:



Any cubic polynomial in e can be written as

$$f^e(x) = \phi_1^e N_1^e(x) + \phi_2^e N_2^e(x) + \phi_3^e N_3^e(x) + \phi_4^e N_4^e(x)\,.$$

Furthermore, it is easy to verify that

$$N_1^e(x_1^e) = 1, \qquad (N_1^e)'(x_1^e) = 0, \qquad N_1^e(x_2^e) = 0, \qquad (N_1^e)'(x_2^e) = 0,$$

$$N_2^e(x_1^e)=0, \quad (N_2^e)'(x_1^e)=1, \quad N_2^e(x_2^e)=0, \quad (N_2^e)'(x_2^e)=0,$$
$$N_3^e(x_1^e)=0, \quad (N_3^e)'(x_1^e)=0, \quad N_3^e(x_2^e)=1, \quad (N_1^e)'(x_2^e)=0,$$
$$N_4^e(x_1^e)=0, \quad (N_4^e)'(x_1^e)=0, \quad N_4^e(x_2^e)=0, \quad (N_4^e)'(x_2^e)=1,$$

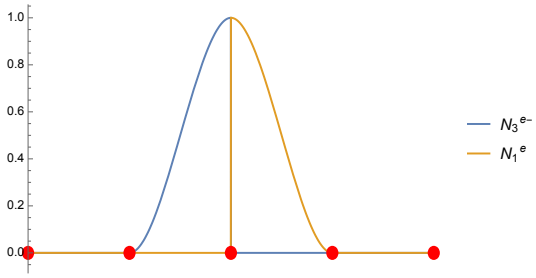
which implies that the **degrees of freedom** are

$$\phi_1^e = f^e(x_1^e), \quad \phi_2^e = (f^e)'(x_1^e), \quad \phi_3^e = f^e(x_2^e), \quad \phi_4^e = (f^e)'(x_2^e), \tag{1.183}$$

which are the values of the function and its derivative at vertices x_1^e and x_2^e. The fact that the degrees of freedom involve not just the value of the function but also the value of its derivative is what qualifies this element as being an **Hermite** finite element. The element then has two nodes, one at each vertex. To pictorially indicate that a degree of freedom at each node is the derivative therein, we draw a ring around the node, i.e.,

These elements can easily be combined in such a way to obtain global basis functions that are $\mathscr{C}^1$ and generate the Hermite space. The basic procedure is as follows:

- The function N_1^e of element e (extended by zero to the rest of the domain) is added to the function N_3^{e-} (also extended by zero), where $e-$ is the element to the left of e (if any). The resulting function over Ω is nothing but the function $H_{2a-1}(x)$, already introduced in (1.164), assuming that a is the *left* node of e.



- For the even-numbered basis functions the construction is analogous. The functions N_2^e and N_4^{e-} are added up. The resulting function is $H_{2a}(x)$, introduced in (1.165), assuming that a is the *left* node of e.

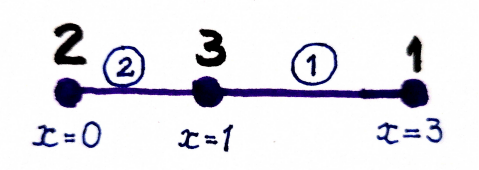
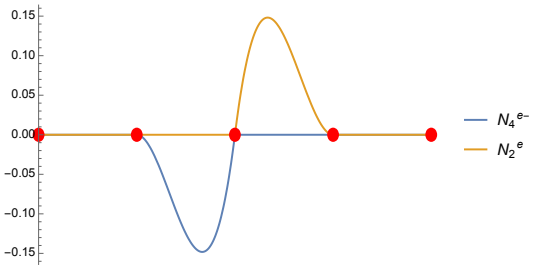


Figure 1.21 Mesh of the example in the text. The circled numbers correspond to the elements, the bare numbers to the nodes.



A key issue is that **these operations can be encoded in a local-to-global map**, so as to use **the same assembly procedure** as in Section 1.4.4. This is best explained by carefully carrying out an example.

Example 1.84

Consider a mesh of just two elements and three nodes. Furthermore, let us adopt an arbitrary numbering of nodes and elements so as to show how general the procedure is. The nodal coordinates are

$$x_1 = 3,\ x_2 = 0,\ x_3 = 1$$

and we specify that element domain number 1 is (x_3, x_1) and element domain number 2 is (x_2, x_3), as shown in Fig. 1.21. Having 3 nodes, the dimension of the cubic Hermite space that we will generate is $m = 2n_{\text{vert}} = 6$.

In this mesh, the following local-to-global map yields a basis of $\mathcal{W}_h$

$$\mathsf{LG} = \begin{pmatrix} 5 & 3 \\ 6 & 4 \\ 1 & 5 \\ 2 & 6 \end{pmatrix}.$$

In fact, from the definition

$$N_A = \mathop{\mathring{\sum}}_{\{(a,e)|\mathsf{LG}(a,e)=A\}} N_a^e$$

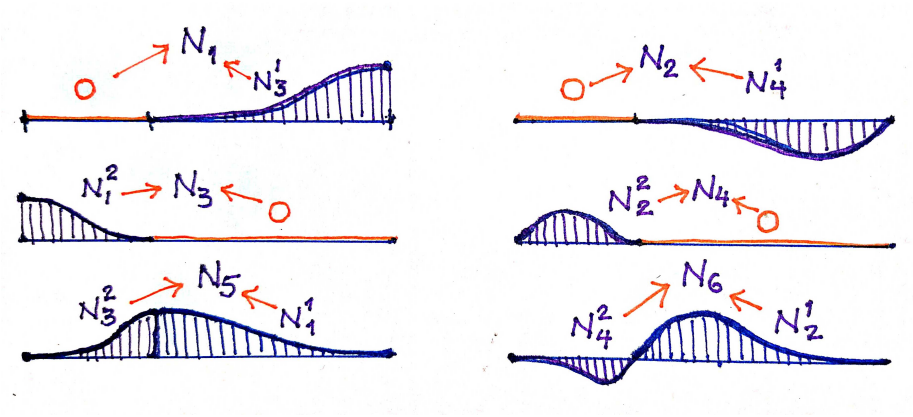


Figure 1.22 The six basis functions generated by the local-to-global map LG provided in the example.

we can write down the sum explicitly for $A = 1,\ldots,6$, yielding

$$\begin{aligned} N_1 &= N_3^1\,, \\ N_2 &= N_4^1\,, \\ N_3 &= N_1^2\,, \\ N_4 &= N_2^2\,, \\ N_5 &= N_1^1 \mathring{+} N_3^2\,, \\ N_6 &= N_2^1 \mathring{+} N_4^2\,. \end{aligned}$$

These functions, together with their elementwise contributions, are shown in Fig. 1.22. They can be directly compared to the Hermite basis functions H_k introduced in (1.164)-(1.165). The local-to-global map has done its magic, since in fact we have

$$N_1 = H_5, \quad N_2 = H_6, \quad N_3 = H_1, \quad N_4 = H_2, \quad N_5 = H_3, \quad N_6 = H_4\,.$$

The Hermite finite element, together with the local-to-global map, have generated the Hermite global basis functions. The numbering is different, but just because we numbered the nodes differently.

Exercise: Verify that with the local-to-global map

$$\mathsf{LG} = \begin{pmatrix} 3 & 1 \\ 4 & 2 \\ 5 & 3 \\ 6 & 4 \end{pmatrix}$$

with the mesh of Fig. 1.21 (without renumbering the elements) produces basis functions that satisfy $N_k = H_k$ for all $k = 1,\ldots,m$.

1.5.6 The element stiffness matrix and load vector

It is clear by now that a mesh of cubic Hermite finite elements can be "assembled," provided a correct local-to-global map is provided, into the piecewice cubic Hermite C^1 space that we denoted H_3-space and introduced as the "simplest" C^1 space and which we have seen to work quite well to solve fourth order problems.

We now look for the element matrices and vectors that will allow us to implement codes in which the mesh is arbitrary and the coefficients are not constant. This will be a significant gain in generality with respect to the method discussed in Example 1.83.

From the bilinear form (1.159b) we know that the contribution of each element is the element stiffness matrix

$$\mathsf{K}^e_{ij} = \int_{x_1^e}^{x_2^e} \left[q\,(N_j^e)"\,(N_i^e)" + c\,N_j^e\,N_i^e \right] dx. \tag{1.184}$$

Both terms in the integrand are products of the (assumed known) functions q and c by polynomials, so that in principle the integral can be computed exactly. We provide next the exact expressions that arise when q and c are **constant within the element**, equal to real numbers q_e and c_e.

The following calculations can be checked by hand:

$$\begin{array}{lcllcl}
\int_{x_1^e}^{x_2^e}(N_1^e)"\,(N_1^e)"\,dx &=& \frac{12}{h_e^3}, & \int_{x_1^e}^{x_2^e} N_1^e\,N_1^e\,dx &=& \frac{13h_e}{35},\\
\int_{x_1^e}^{x_2^e}(N_2^e)"\,(N_1^e)"\,dx &=& \frac{6}{h_e^2}, & \int_{x_1^e}^{x_2^e} N_2^e\,N_1^e\,dx &=& \frac{11h_e^2}{210},\\
\int_{x_1^e}^{x_2^e}(N_3^e)"\,(N_1^e)"\,dx &=& -\frac{12}{h_e^3}, & \int_{x_1^e}^{x_2^e} N_3^e\,N_1^e\,dx &=& \frac{9h_e}{70},\\
\int_{x_1^e}^{x_2^e}(N_4^e)"\,(N_1^e)"\,dx &=& \frac{6}{h_e^2}, & \int_{x_1^e}^{x_2^e} N_4^e\,N_1^e\,dx &=& -\frac{13h_e^2}{420},\\
\int_{x_1^e}^{x_2^e}(N_2^e)"\,(N_2^e)"\,dx &=& \frac{4}{h_e}, & \int_{x_1^e}^{x_2^e} N_2^e\,N_2^e\,dx &=& \frac{h_e^3}{105},\\
\int_{x_1^e}^{x_2^e}(N_3^e)"\,(N_2^e)"\,dx &=& -\frac{6}{h_e^2}, & \int_{x_1^e}^{x_2^e} N_3^e\,N_2^e\,dx &=& \frac{13h_e^2}{420},\\
\int_{x_1^e}^{x_2^e}(N_4^e)"\,(N_2^e)"\,dx &=& \frac{2}{h_e}, & \int_{x_1^e}^{x_2^e} N_4^e\,N_2^e\,dx &=& -\frac{h_e^3}{140},\\
\int_{x_1^e}^{x_2^e}(N_3^e)"\,(N_3^e)"\,dx &=& \frac{12}{h_e^3}, & \int_{x_1^e}^{x_2^e} N_3^e\,N_3^e\,dx &=& \frac{13h_e}{35},\\
\int_{x_1^e}^{x_2^e}(N_4^e)"\,(N_3^e)"\,dx &=& -\frac{6}{h_e^2}, & \int_{x_1^e}^{x_2^e} N_4^e\,N_3^e\,dx &=& -\frac{11h_e^2}{210},\\
\int_{x_1^e}^{x_2^e}(N_4^e)"\,(N_4^e)"\,dx &=& \frac{4}{h_e}, & \int_{x_1^e}^{x_2^e} N_4^e\,N_4^e\,dx &=& \frac{h_e^3}{105},
\end{array}$$

where $h_e = x_2^e - x_1^e$. Then the (symmetric) element stiffness matrix ends up being

$$\mathsf{K}^e = \begin{pmatrix}
\frac{12q_e}{h_e^3} + \frac{13c_eh_e}{35} & \frac{6q_e}{h_e^2} + \frac{11c_eh_e^2}{210} & -\frac{12q_e}{h_e^3} + \frac{9c_eh_e}{70} & \frac{6q_e}{h_e^2} - \frac{13c_eh_e^2}{420}\\
\text{symm} & \frac{4q_e}{h_e} + \frac{c_eh_e^3}{105} & -\frac{6q_e}{h_e^2} + \frac{13c_eh_e^2}{420} & \frac{2q_e}{h_e} - \frac{c_eh_e^3}{140}\\
\text{symm} & \text{symm} & \frac{12q_e}{h_e^3} + \frac{13c_eh_e}{35} & -\frac{6q_e}{h_e^2} - \frac{11c_eh_e^2}{210}\\
\text{symm} & \text{symm} & \text{symm} & \frac{4q_e}{h_e} + \frac{c_eh_e^3}{105}
\end{pmatrix} \tag{1.185}$$

Turning now to the element load vector, we will compute it without the end contributions, which will be added later on. From (1.159c) we have

$$\mathsf{F}_i^e = \int_{x_1^e}^{x_2^e} f(x)\, N_i^e(x)\, dx, \tag{1.186}$$

which again can in principle computed exactly.

As an interesting special case, let us compute F^e explicitly for the case in which $f(x) = f_e$, constant within the element. This will allows us to solve problems with piecewise-constant distributed load.

From the straightforward integrals

$$\int_{x_1^e}^{x_2^e} N_1^e(x)\, dx = \int_{x_1^e}^{x_2^e} N_3^e(x)\, dx = \frac{h_e}{2}\,, \qquad \int_{x_1^e}^{x_2^e} N_2^e(x)\, dx = -\int_{x_1^e}^{x_2^e} N_4^e(x)\, dx = \frac{h_e^2}{12}\,,$$

we get the required expression:

$$\mathsf{F}^e = \begin{pmatrix} \frac{f_e h_e}{2} \\ \\ \frac{f_e h_e^2}{12} \\ \\ \frac{f_e h_e}{2} \\ \\ -\frac{f_e h_e^2}{12} \end{pmatrix} \tag{1.187}$$

These expressions can be coded in the element routine:

```
function [Ke, Fe]=elementKandF(xe,qe,ce,fe)
  he=xe(2)-xe(1);
  qh=qe/he;qh2=qh/he;qh3=qh2/he;
  ch=ce*he;ch2=ch*he;ch3=ch2*he;
  Ke=[12*qh3+13*ch/35, 6*qh2+11*ch2/210, -12*qh3+9*ch/70, 6*qh2-13*ch2/420;...
      6*qh2+11*ch2/210, 4*qh+ch3/105, -6*qh2+13*ch2/420, 2*qh-ch3/140;...
      -12*qh3+9*ch/70, -6*qh2+13*ch2/420, 12*qh3+13*ch/35, -6*qh2-11*ch2/210;...
      6*qh2-13*ch2/420, 2*qh-ch3/140, -6*qh2-11*ch2/210, 4*qh+ch3/105];
  fh=fe*he;fh2=fh*he;
  Fe=[fh/2; fh2/12; fh/2; -fh2/12];
end
```

1.5.7 Solving Fourth-order Elliptic Problems with H_3 Hermite Finite Elements

We assume that a mesh of H_3 finite elements is provided by means of a **list of coordinates** X and a **local-to-global map** LG.

The specified values g_0, d_0, T and F are also provided, together with the piecewise constant values for q_e, c_e and f_e.

We are thus in a position to code the assembly of the stiffness matrix and load vector. We follow the same procedure as in the case of P_1 elements. Notice

that, as before, we are looking for an array $\mathsf{U} = (u_1, u_2, \ldots, u_{2n_{\text{nod}}})^T$ that defines the solution of the variational method as

$$u_h(x) = u_1 N_1(x) + u_2 N_2(x) + \ldots + u_{2n_{\text{nod}}} N_{2n_{\text{nod}}}(x)$$

where we used $n_{\text{nod}} = n_{\text{vert}}$, since the number of nodes is equal to the number of vertices in this case.

The code starts by identifying n_{nod}, n_{el} and m from the data, and initializing K and F to zero.

```
nod=length(X); nunk=2*nod; nel=size(LG,2);
K=zeros(nunk,nunk);F=zeros(nunk,1);
```

Then, assuming that the elementwise values of $q(x)$, $c(x)$ and $f(x)$ are stored in the arrays qq, cc and ff, respectively, we proceed to assemble the contributions of the element stiffness matrices and load vectors.

```
for iel=1:nel
%% setting the local data
  lge=LG(:,iel);
  xe(1,1:npe)=X(1,iel:iel+1);
  qe=qq(iel);   ce=cc(iel);  fe=ff(iel);
%% computing element K and F
  [Ke Fe]=elementKandF(xe,qe,ce,fe);
%% assembly, from local to global
  for ii=1:4
     if (sum(EtaG==lge(ii))==0)
        for jj=1:4
            K(lge(ii),lge(jj))=K(lge(ii),lge(jj))+Ke(ii,jj);
        end
        F(lge(ii))=F(lge(ii))+Fe(ii);
     end
  end
end
```

Notice that this procedure is exactly the same as that used for all other finite element spaces.

Finally, we impose the **essential boundary conditions**, i.e., the specified values (array GG) of the unknowns listed in η_g (array EtaG)

```
ng=length(EtaG);
for ig=1:ng
 K(EtaG(ig),EtaG(ig))=1;
 F(EtaG(ig))=GG(ig);
end
```

and the **natural boundary conditions**, i.e., the torque T and the force F

```
% adding natural boundary conditions at last two unknowns
F(nunk-1)=F(nunk-1)+FL;
F(nunk)=F(nunk)+TL;
```

With this, we can compute the coefficients $u_1, u_2, \ldots, u_{2n_{\text{nod}}}$ by solving the linear system $\mathsf{KU} = \mathsf{F}$.

```
%% solve algebraic system
U=K\F;
```

Omitting the input and output sections of the code, the whole finite element procedure consists of about 40 lines of Octave/MATLAB code.

Example 1.85 (The Euler-Bernoulli beam equation with non-constant bending rigidity and non-uniform mesh)

Consider as example a beam with the same parameters as that plotted in Fig. 1.20, but now we will use a non-uniform mesh with the following array of coordinates

$$\mathsf{X} = \begin{pmatrix} 0 & 0.25 & 0.4 & 0.6 & 0.65 & 0.9 & 1 \end{pmatrix},$$

so that $n_{\text{nod}} = 7$, $n_{\text{el}} = 6$, and a suitable local-to-global map, such as the natural one

$$\mathsf{LG} = \begin{pmatrix} 1 & 3 & 5 & 7 & 9 & 11 \\ 2 & 4 & 6 & 8 & 10 & 12 \\ 3 & 5 & 7 & 9 & 11 & 13 \\ 4 & 6 & 8 & 10 & 12 & 14 \end{pmatrix}.$$

Because of the boundary conditions we have $\eta_g = \{1,2\}$ with $\mathsf{GG} = (0,0)$, and natural boundary conditions on the right $T = 0$ and $F = 0$.

Running the code one obtains the solution shown in Fig. 1.23, which is very similar to that of Fig. 1.20 up to small discretization errors. But notice that, in this case, the mesh is non-uniform.

Furthermore, we can vary the values of the bending rigidity of each element at will. Notice that element number 4 is quite small, it goes from $x = 0.6$ to $x = 0.65$. Let us change the bending stiffness of just this element to 1/100-th that of the rest of the beam, from $q_e = 10$ to $q_e = 0.1$. The solution radically changes, as seen in Fig. 1.24. Element number 4, being less stiff, acts as a hinge at which most of the deformation concentrates.

Example 1.86 (Image denoising) We can apply the same method and code as before to an image denoising problem by solving the equation

$$q\,u''''(x) + u(x) = u_0(x),$$

where u_0 is the raw image and q an adjustable parameter, with homogeneous natural boundary conditions ($T = F = 0$ at both ends). We use the H_3-space with one element per pixel in the image. The results are shown in Figure 1.25. The denoising effect of the fourth-order term is evident. A value $q = 10^{-6}$ seems the most appropriate for this image. For $q > 10^{-5}$ the solution is too smoothed, loosing the underlying signal. For $q < 10^{-7}$ the solution follows the local noise, which is not sufficiently removed.

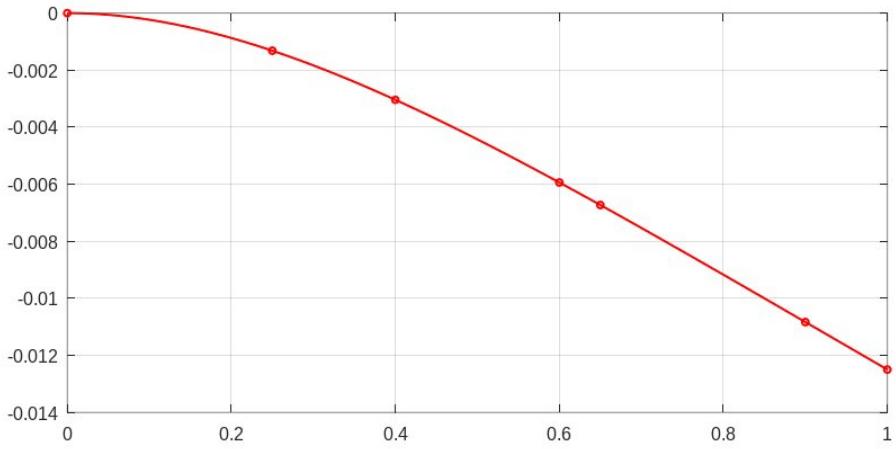


Figure 1.23 Finite element solution from Example 1.85.

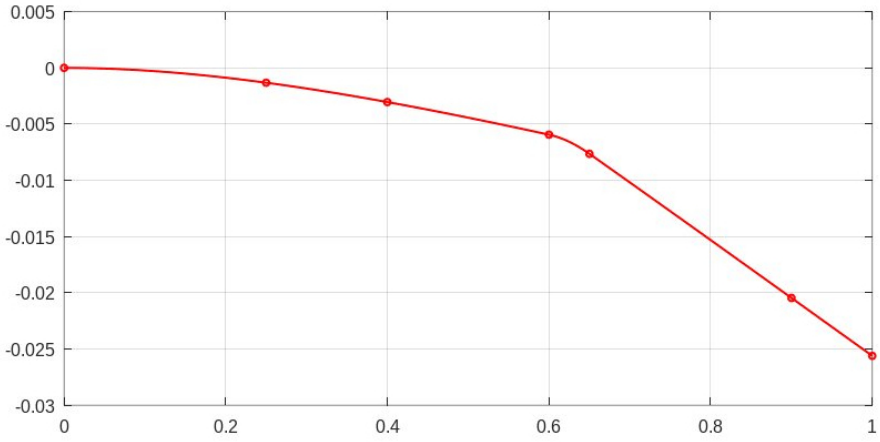


Figure 1.24 Finite element solution from Example 1.85. The bending rigidity of the beam is equal to 10 in all elements except for element 4, where its value has been reduced to 0.1.

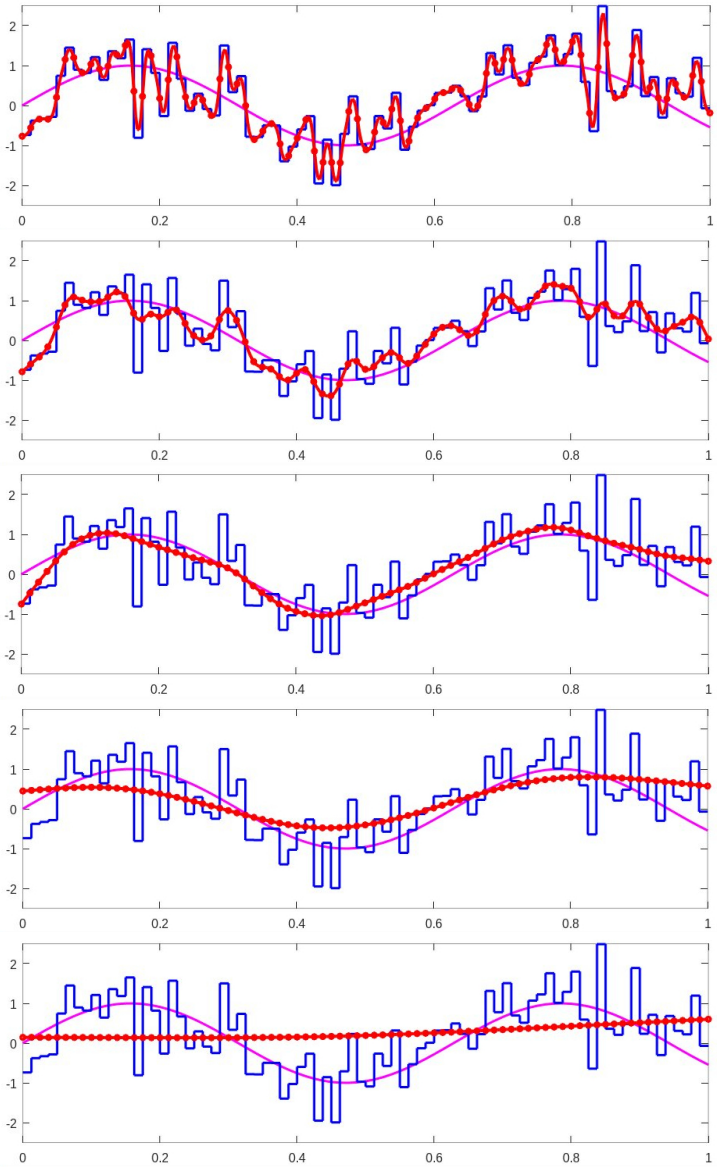


Figure 1.25 Results from the image denoising example. In blue the original image u_0, which is a random perturbation of the exact function drawn in magenta. In red we plot the solution u_h of the image-denoising example for q taking values, from top to bottom, $q = 10^{-10}$, 10^{-8}, 10^{-6}, 10^{-4} and 10^{-2}.

Chapter 2

Diffusion Problems in 2D

We now turn to consider diffusion problems in two dimensions, which govern heat conduction in solids, electrostatics and some mass transfer situations. We follow the same methodology as in Chapter 1, beginning with the partial differential equation of the mathematical problem and deriving a variational equation in the standard way, discussing the novelties brought by the higher dimensionality of the domain, and finally introducing the simplest finite element method to compute an approximate solution.

2.1 The Partial Differential Equation

Consider a general **diffusion equation**

$$-\operatorname{div}(K\nabla u) = f\,, \tag{2.1}$$

which should be satisfied by a function $u\colon \Omega \to \mathbb{R}$ in a domain $\Omega \subset \mathbb{R}^2$. It is convenient to denote the Cartesian coordinates by $x = (x_1, x_2)^T$, and the partial derivatives of a function $u(x_1, x_2)$ by $\partial_1 u$ and $\partial_2 u$. The components of a vector v in a basis to be specified will be likewise denoted by $(v_1, v_2)^T$.

Above, div is the divergence operator which applied to a vector field $v\colon \Omega \to \mathbb{R}^2$ yields

$$\operatorname{div} v = \partial_1 v_1 + \partial_2 v_2,$$

K is a positive-definite symmetric matrix (all eigenvalues are positive), ∇u is the gradient vector

$$\nabla u = (\partial_1 u\,, \partial_2 u\,)^T$$

and $f\colon \Omega \to \mathbb{R}$ is a source density (per unit area). When u represents the temperature of a solid, (2.1) is known as the **heat conduction equation**.

The diffusion equation (2.1) can be recast as $\operatorname{div} J = f$, where

$$J = -K\nabla u\,, \tag{2.2}$$

is the **diffusive flux vector** (or **heat flux vector** in the thermal setting). In the context of the heat conduction equation, relationship (2.2) is called **Fourier's law**.

In the context of mass transport, u is the concentration of mass and (2.2) is called **Fick's law**.

All the previous expressions have been written in operator form, which is a concise way of writing formulae involving partial derivatives. They can of course be rewritten in (Cartesian) coordinates, as

$$-\sum_{i=1}^{2}\partial_i\left(\sum_{j=1}^{2}K_{ij}\partial_j u\right) = f\,, \tag{2.3}$$

$$J_i = -\sum_{j=1}^{2}K_{ij}\partial_j u. \tag{2.4}$$

Notice that we have not adopted notation that indicates if a symbol is a scalar, a vector, or a matrix; the nature of a symbol will be interpreted from the context. For additional background material on the divergence operator, including its definition for domains in three dimensions, we refer the reader to [7, Ch. 2].

Example 2.1 (The Poisson equation) When K is a multiple of the identity matrix,

$$K(x) = k(x)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},$$

we say that the diffusive medium is **isotropic**. If we further assume that k is independent of x we have the case in which (2.1) is a **Poisson equation**. In such a case the expressions simplify considerably. Since $K_{ij} = k\delta_{ij}$, we have

$$K = \begin{pmatrix}k & 0\\ 0 & k\end{pmatrix}, \tag{2.5}$$

$$J = -K\nabla u = -k\nabla u = \begin{pmatrix}-k\partial_1 u\\ -k\partial_2 u\end{pmatrix}, \tag{2.6}$$

and thus (2.1) reads

$$\partial_{11}^2 u + \partial_{22}^2 u = -\frac{f}{k}\,. \tag{2.7}$$

The notation of the second partial derivatives is

$$\partial_{ij}^2 u = \frac{\partial^2 u}{\partial x_i \partial x_j}$$

and one can recognize in the left-hand side of (2.7) the **Laplacian** of u, namely

$$\Delta u = \left(\partial_{11}^2 + \partial_{22}^2\right)u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\,. \tag{2.8}$$

There exist many analytical solutions of $\Delta u = -f/k$ even if $f = c$ is a constant. In the case $c = 0$ the following functions are solutions for any α, β and γ real numbers,

$$u(x_1, x_2) = \alpha + \beta x_1 + \gamma x_2\,. \tag{2.9}$$

In other words, all *affine functions* are solutions to $\Delta u = 0$. This is also the case in 1D, where all functions of the form $u(x) = \alpha + \beta x$ are solutions to $u'' = 0$. There is a crucial difference though: In 1D the affine functions are *all possible solutions* to $u'' = 0$, while in 2D there are infinitely many linearly independent functions that satisfy $\Delta u = 0$, also called **harmonic functions**. For example, the function

$$u(x_1, x_2) = \ln\left((x_1 - X_1)^2 + (x_2 - X_2)^2\right) \tag{2.10}$$

defined for all $x \neq X$ satisfies $\Delta u = 0$ in $\mathbb{R}^2 \setminus X$, for *any choice of* X. These functions are smooth in Ω if $X \notin \overline{\Omega}$.[1] Two such functions with different choices of X are linearly independent, so that the set of harmonic functions has infinite dimensions.

Example 2.2 (The Elastic Membrane) The diffusion equation (2.1) also appears when modeling small deformations of planar elastic membranes under tension and subjected to loads normal to the undeformed surface [11, §93], see Fig. 2.1. Specifically, in this case the membrane occupies a domain $\Omega \subset \mathbb{R}^2$, and the unknown function u is the vertical displacement normal to the membrane. The membrane is "pulled" with a force per unit length of magnitude T normal to its boundary, and pressurized with a pressure p along the vertical direction, while the boundary or part of the boundary $\partial\Omega$ is prevented from moving. The problem consists in determining the deformed shape of the membrane $(x, y, u(x, y)) \in \mathbb{R}^3$, where $(x, y) \in \Omega$. The equation that defines u is

$$p = -\operatorname{div}(T\nabla u), \tag{2.11}$$

Example 2.3 (Torsion of a Prismatic Bar) Consider a prismatic bar B of cross-section Ω and length L, so that the bar occupies the set of points $B = \Omega \times [0, L]$ in $\mathbb{R}^3$, see Fig. 2.2. The top surface of the bar, $\Omega \times \{L\}$, is rotated by an angle θL relative to the bottom surface $\Omega \times \{0\}$. The bar is made of a linear elastic material, and the shear stress components (τ_1, τ_2) on any cross section $\Omega \times \{x_3\}$, $x_3 \in [0, L]$, is computed from the so-called stress function $\phi\colon \Omega \to \mathbb{R}$ [11, §90], which does not depend on x_3, and satisfies the Poisson equation

$$\frac{\partial^2 \phi}{\partial x_1^2} + \frac{\partial^2 \phi}{\partial x_2^2} = -2\mu\theta$$

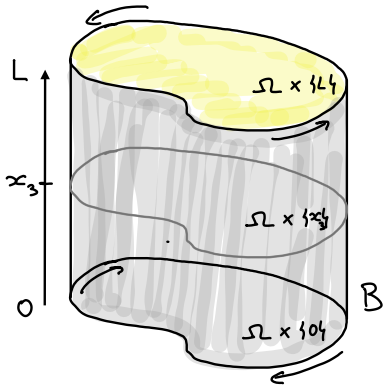


Figure 2.2 Torsion of a prismatic bar B with cross section Ω.

[1] The set $\overline{\Omega}$ is the **closure** of Ω, or the set of points in Ω and all those points that can be reached as limits of sequences of points in Ω. We have seen this concept for one-dimensional domains in §1, and in the discussion on the sparsity of the stiffness matrix in §1.4.4.

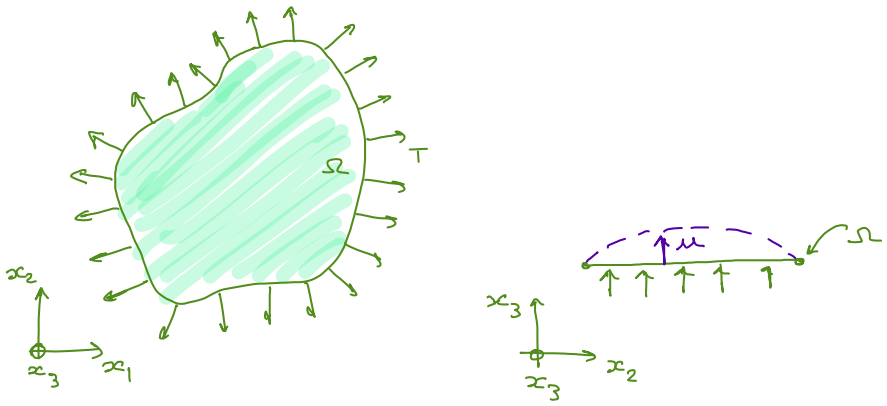


Figure 2.1 A planar elastic membrane under an isotropic tension T, pressurized from the bottom. The shape of the elastic membrane is shown on the left, as seen from above, while a view parallel to the membrane, on the right, shows the pressure loading p at the bottom, and the vertical displacement u of each point in the membrane that defines its deformed shape.

in Ω, where μ is the shear modulus. The shear stress components follow from ϕ as

$$\begin{bmatrix}\tau_x\\ \tau_y\end{bmatrix} = \begin{bmatrix}\partial_2\phi\\ -\partial_1\phi\end{bmatrix}.$$

Two-dimensional domains. The diffusion equation (2.1) is assumed to hold at all points x of the **two-dimensional domain** Ω. In 1D the domains could be intervals, or at most groups of intervals. The diversity of domains in 2D is much larger. The shape of the domain usually comes from the geometry of the physical system under study. The theory and methods we describe below hold for **bounded** domains that do not have cusps or cracks. A visual guide of the **admissible domains** is given in Fig. 2.3. To avoid unnecessary technical discussions at this stage of the learning, however, we restrict our attention to **polygonal domains**, with the possibility of them having one or several polygonal holes. The **boundary** of a domain Ω, denoted $\partial\Omega$, is formed by one or more polygonal lines. We will have an opportunity to consider curved domains later.

Boundary conditions. Boundary conditions are necessary to uniquely identify a solution of (2.1). A salient feature of elliptic second-order problems is that at all points in $\partial\Omega$ one (and only one) piece of information is needed about the solution u. We assume that $\partial\Omega$ is decomposed into two parts, depending on the available boundary information.

- If at $x \in \partial\Omega$ we know the **value** of u, we say that x belongs to the **Dirichlet boundary** $\partial\Omega_D$.
- On the other hand, if at x we know the value of the **normal flux** $J\cdot\check{n}$, with $\check{n}$ the **exterior unit normal to** $\partial\Omega$ at x, we say that x belongs to the **Neumann boundary** $\partial\Omega_N$.

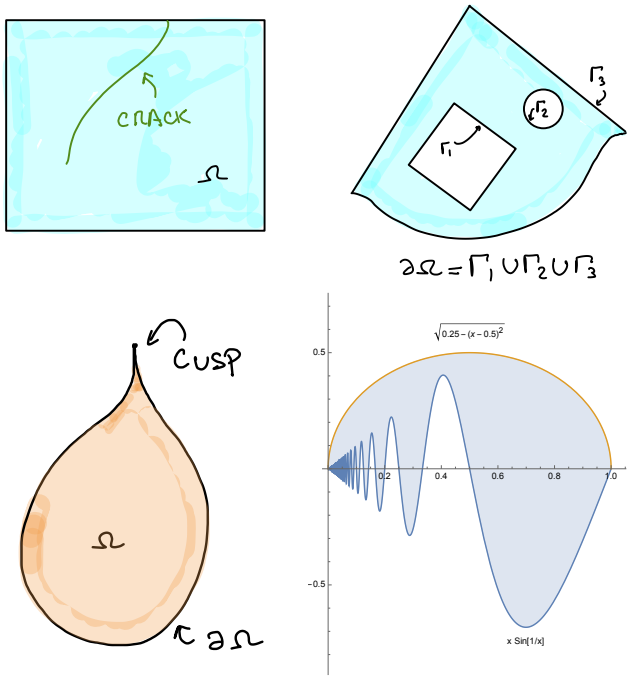


Figure 2.3 Domains that we will consider can contain smooth cracks, holes, and have a boundary made of a finite number of smooth curves, each of finite length, such as polygonal domains, as shown in the top row. Domains that we will not consider have boundaries that form cusps (vertices with the same tangent line on each side), have boundaries that have infinite length, or that do not have normal to the boundary defined at almost every point of the boundary (e.g., by replacing the function $x\sin(1/x)$ for a fractal curve, such as the Weierstrass function [5]), as illustrated in the bottom row. Typical engineering domains are idealized to be of the type in the top row.

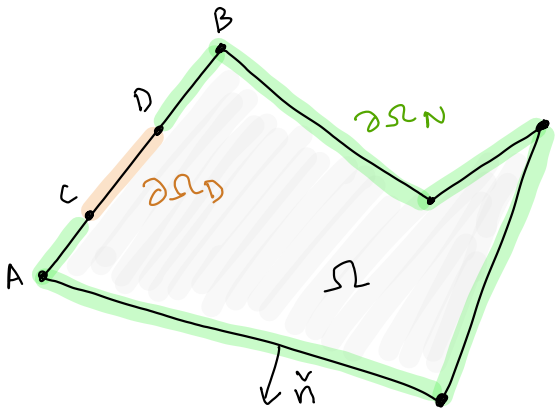


Figure 2.4 Sketch of the domain for Problem 2.1

- There exist other possibilities, such as **Robin boundary conditions**, in which we know the value of a linear combination of the function and the normal flux.

We assume that both $\partial\Omega_D$ and $\partial\Omega_N$ consist of a subset of edges of $\partial\Omega$, which is a polygonal line (or several).

The problem whose solution we aim to approximate reads as follows.

Problem 2.1 (Strong Form of the 2D Diffusion Problem)**.** *Given the coefficients K and f as functions of $x \in \Omega$, and given a real function g defined in $\partial\Omega_D$ and another real function H defined in $\partial\Omega_N$, determine a function $u : \Omega \to \mathbb{R}$ satisfying*

$$-div\,(K\nabla u) = f(x) \qquad \forall\, x \in \Omega \tag{2.12a}$$

$$u = g \qquad \forall\, x \in \partial\Omega_D \tag{2.12b}$$

$$(K\nabla u)\cdot \check{n} = H \qquad \forall\, x \in \partial\Omega_N \tag{2.12c}$$

Problem 2.1 admits one and only one solution under sufficient smoothness of the data (in particular, g must be continuous) plus the two essential hypotheses:

H1) the thermal conductivity K is everywhere a bounded, positive definite matrix, with all eigenvalues greater than some $\kappa_0 > 0$, and

H2) the length of $\partial\Omega_D$ is strictly positive.

If K is a multiple of the identity matrix, i.e., $K(x) = k(x)\mathbf{I}_{2\times 2}$, then H1 requires that $k(x) > \kappa_0 > 0$ for all $x \in \Omega$. Concerning H2, it requires that u is known not just at a point or a finite set of points of $\partial\Omega$, the condition $u = g$ must hold all along the full length of an edge of $\partial\Omega$. Notice that if $\partial\Omega_D$ is just a segment $\overline{CD}$ within a larger edge $\overline{AB}$, it is possible to redefine the polygon incorporating C and D as vertices, so that $\partial\Omega_D$ is a full edge, c.f. Fig. 2.4.

Example 2.4 (A uniformly heated rod) Consider the circular cross section of a homogeneous and isotropic rod, in which heat is generated uniformly at rate f. The domain is thus $\Omega = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 < R^2\}$.

Let $g \in \mathbb{R}$ be the temperature at the rod's surface, assumed uniform. We are interested in the temperature field **inside** the rod.

The corresponding differential equation is

$$\Delta u = -\frac{f}{k},$$

so, we are looking for $u(x, y)$, satisfying $u = g$ on the circle $x_1^2 + x_2^2 = R^2$ and having (constant) Laplacian equal to $-f/k$ in the enclosed region. The Dirichlet boundary $\partial\Omega_D$ is the whole boundary $\partial\Omega$ of the domain, and thus $\partial\Omega_N$ is empty. The constant k is assumed positive, so that both H1 and H2 are satisfied.

It is easy to check that

$$u(x, y) = g - \frac{f}{4k}\left(x_1^2 + x_2^2 - R^2\right)$$

satisfies these conditions and is thus the unique solution to the boundary value problem.

By differentiating u we can compute the heat flux

$$J = -k\nabla u = -k\left(-\frac{f x_1}{2k}, -\frac{f x_2}{2k}\right)^T = \frac{f}{2}x$$

and see that it is constant along the boundary circle, pointing outwards and with magnitude $fR/2$.

The maximum temperature takes place at the center (if $f > 0$), with value $g + fR^2/(4k)$.

Important: The same solution $u(x_1, x_2)$ also satisfies the **Neumann** boundary value problem, in which the **normal flux** is specified as

$$(k\nabla u) \cdot \check{n} = -\frac{fR}{2}$$

over the boundary circle. However, this problem **does not satisfy H2** (because $\partial\Omega_D$ is empty) and in facts admits not just u as solution but also any function $v = u + C$, with C an arbitrary real constant.

2.2 A Variational Equation

As already discussed in the 1D case, the finite element method is built upon a *variational equation* for the problem under consideration, which at present is Problem 2.1. Getting to a variational formulation often involves integration by parts, so let us recall a useful result.

Theorem 2.1. (Integration by parts in 2D or 3D) *Let w be a smooth vector field in Ω (an admissible domain), and v a smooth scalar function. Then,*

$$\int_\Omega v \operatorname{div} w \, d\Omega = \int_{\partial\Omega} v \, w \cdot \check{n} \, d\Gamma - \int_\Omega w \cdot \nabla v \, d\Omega \,. \tag{2.13}$$

It is also convenient to illustrate how the divergence theorem, (2.13) is written in Cartesian components in $\mathbb{R}^d$, where d is the dimension of the space. The components of w are $(w_1, \ldots, w_d)$. Then:

$$\sum_{i=1}^{d} \left[\int_\Omega v \partial_i w_i \, d\Omega \right] = \sum_{i=1}^{d} \left[\int_{\partial\Omega} v w_i \check{n}_i \, d\Gamma - \int_\Omega w_i \partial_i v \, d\Omega. \right] \tag{2.14}$$

Then, a useful mnemonic rule to apply integration by parts in higher dimensions is to move the "index" i from indicating a partial derivative of the component w_i to the partial derivative of the function v, or viceversa.

Example 2.5 Let's consider an example of the use of the theorem. Let $\Omega = [0,1]^3$ and $w(x) = x_i e_i$, where x_i are the components of the point x in a Cartesian basis e_i, and $v(x) = 1$, a constant function. Then, div $w = 3$, and hence

$$\int_\Omega \operatorname{div} w \, dV = 3. \tag{2.15}$$

Alternatively, let's compute

$$\int_{\partial\Omega} w \cdot \check{n} \, d\Gamma - \int_\Omega w \cdot \nabla v \, d\Omega. \tag{2.16}$$

We have that in the face where $x_1 = 0$

$$\int_{\{x_1=0\}\cap\partial\Omega} w \cdot \check{n} \, d\Gamma = \int_{\{x_1=0\}\cap\partial\Omega} (x_2 e_2 + x_3 e_3) \cdot (-e_1) \, d\Gamma = 0, \tag{2.17}$$

and the same happens in the faces defined by $x_2 = 0$ and $x_3 = 0$. Alternatively, when $x_1 = 1$, we have

$$\int_{\{x_1=1\}\cap\partial\Omega} w \cdot \check{n} \, d\Gamma = \int_{\{x_1=0\}\cap\partial\Omega} (e_1 + x_2 e_2 + x_3 e_3) \cdot e_1 \, d\Gamma = 1, \tag{2.18}$$

and the same thing happens in the faces defined by $x_2 = 1$ and $x_3 = 1$. Since $\nabla v = 0$, we verified that

$$\int_{\partial\Omega} w \cdot \check{n} \, d\Gamma - \int_\Omega w \cdot \nabla v \, d\Omega = 3, \tag{2.19}$$

as the theorem states.

Now, applying the same recipe as in 1D, §1.1.2.3, we multiply the differential equation (2.12a) by a smooth $v : \Omega \to \mathbb{R}$ and integrate over Ω, to get

$$-\int_\Omega \operatorname{div}(K\nabla u)\, v \, d\Omega = \int_\Omega f\, v \, d\Omega \,.$$

Using (2.13) with $w = K\nabla u$ for the left-hand side and decomposing the integral over $\partial\Omega$ into the sum of $\int_{\partial\Omega_D}$ and $\int_{\partial\Omega_N}$, we arrive to

$$\int_\Omega (K\nabla u)\cdot\nabla v \, d\Omega = \int_\Omega f\, v \, d\Omega + \int_{\partial\Omega_N} v\, K\nabla u\cdot\check{n}\, d\Gamma + \int_{\partial\Omega_D} v\, K\nabla u\cdot\check{n}\, d\Gamma \,. \quad (2.20)$$

In the integral over $\partial\Omega_N$, $K\nabla u\cdot\check{n}$ can be replaced by H, since (2.12c) holds. This makes Neumann boundary conditions to be **natural** boundary conditions. On the other hand, in the integral over $\partial\Omega_D$ we have no way to know $K\nabla u\cdot\check{n}$, so we require $v = 0$ there. The **test space** is then given by

$$\mathcal{V} = \{v\colon \Omega\to\mathbb{R} \text{ smooth } \mid v(x) = 0 \text{ for all } x\in\partial\Omega_D\} \,, \quad (2.21a)$$

so that the last integral in (2.20) is zero.

The solution u, which is assumed to be smooth, thus satisfies the following variational equation:

$$a(u,v) = \ell(v) \qquad \forall\, v\in\mathcal{V} \,, \quad (2.21b)$$

where the bilinear and linear forms are given by

$$a(u,v) = \int_\Omega (K\nabla u)\cdot\nabla v \, d\Omega \,, \quad (2.21c)$$

$$\ell(v) = \int_\Omega f\, v\, d\Omega + \int_{\partial\Omega_N} H\, v\, d\Gamma \,. \quad (2.21d)$$

For completeness, the weak form of the problem reads

Problem 2.2. *(Weak Form of the 2D Diffusion Problem) Let*

$$\mathcal{S} = \left\{v\colon \Omega\to\mathbb{R} \textit{ smooth } \mid v(x) = g(x) \textit{ for all } x\in\partial\Omega_D\right\} \,. \quad (2.22)$$

Find $u\in\mathcal{S}$ such that $a(u,v) = \ell(v)$ for all $v\in\mathcal{V}$.

As you can see, the variational equation involves just first-order derivatives (the gradient ∇), but instead of having the pointwise requirement that $\operatorname{div}(K\nabla u) + f = 0$ at all points we have integral expressions in two dimensions that must hold for all $v\in\mathcal{V}$.

2.2.1 Other Variational Equations

Just as we did in §1.1.2.4, we can obtain a new variational equation that the solution u satisfies through linear combinations of other variational equations it satisfies. Let's see this in this case.

Example 2.6 Nitsche's Method. The solution u of Problem 2.1 satisfies variational equation (2.20) for any $v \in \mathcal{V} = \{w\colon \Omega \to \mathbb{R} \text{ smooth}\}$, and after replacing $K\nabla u \cdot \check{n}$ by H, it reads

$$\int_\Omega (K\nabla u)\cdot\nabla v\, d\Omega - \int_{\partial\Omega_D} vK\nabla u\cdot\check{n}\, d\Gamma = \int_\Omega fv\, d\Omega + \int_{\partial\Omega_N} vH\, d\Gamma. \quad (2.23a)$$

Additionally, u also satisfies the following variational equations

$$\int_{\partial\Omega_D} (g-u)K\nabla v\cdot\check{n}\, d\Gamma = 0 \quad (2.23b)$$

$$\int_{\partial\Omega_D} \mu(u-g)v\, d\Gamma = 0. \quad (2.23c)$$

for all $v \in \mathcal{V}$, where $\mu > 0$ is a positive real number. Adding the three equations in (2.23), we obtain the following variational equation that u also satisfies

$$\begin{aligned}\int_\Omega (K\nabla u)\cdot\nabla v\, d\Omega - \int_{\partial\Omega_D} (vK\nabla u + uK\nabla v)\cdot\check{n}\, d\Gamma + \int_{\partial\Omega_D} \mu uv\, d\Gamma \\ = \int_\Omega fv\, d\Omega + \int_{\partial\Omega_N} vH\, d\Gamma - \int_{\partial\Omega_D} gK\nabla v\cdot\check{n}\, d\Gamma + \int_{\partial\Omega_D} \mu gv\, d\Gamma \quad (2.24)\end{aligned}$$

for all $v \in \mathcal{V}$. Computing the Euler-Lagrange equations would show that both the Dirichlet and Neumann boundary conditions are natural boundary conditions for this variational equation.

2.3 Variational Numerical Methods

As in the 1D case, a linear variational method for finding an approximation to the exact solution u consists of finding u_h that is the solution to Problem 1.2. There are, however, significant differences with the one-dimensional case:

- The spaces $\mathcal{S}_h$ and $\mathcal{V}_h$ must now be composed of functions that take values over a two-dimensional domain.
- The domain boundary is a closed line, which we assumed to be a polygon for simplicity. The construction of the basis of $\mathcal{V}_h$ must ensure that the functions are zero on the boundary $\partial\Omega_D$.
- As we will have the chance to see, a variational method based on variational equation 2.21 is *consistent* if functions in the test space $\mathcal{V}_h$ are continuous. Because $\mathcal{V}_h$ is the direction of $\mathcal{S}_h$, we will set $\mathcal{S}_h$ and $\mathcal{W}_h$ to be spaces of continuous functions. So, if the domain is subdivided into element domains, continuity of the global basis functions between elements must be enforced. This continuity, contrary to the 1D case, must hold not just at the nodes but along all edges of the subdivision.

All functions in $\mathcal{S}_h$ are obtained as $\overline{u}_h + v_h$ for all $v_h \in \mathcal{V}_h$. If all functions in $\mathcal{V}_h$ are continuous, and $\overline{u}_h$ is discontinuous somewhere, then *all* functions in $\mathcal{S}_h$ are discontinuous. While it would be possible to construct approximations to the exact solution u in this way, it is not necessary nor convenient at this stage.

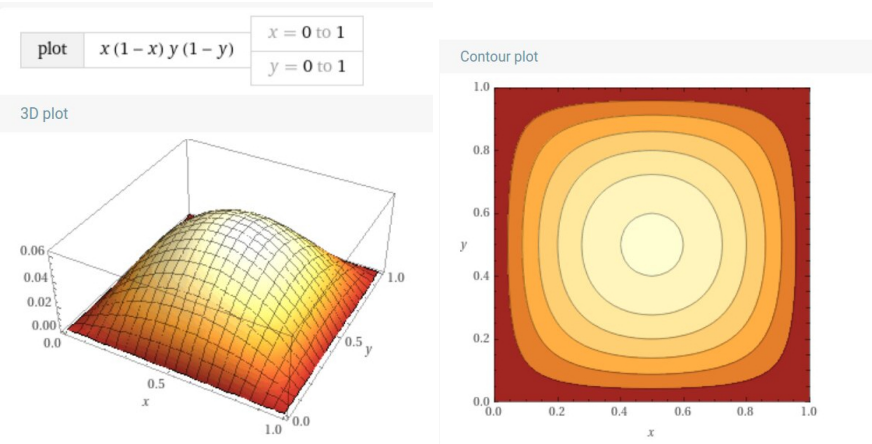


Figure 2.5 The function $N_1(x_1, x_2)$ when $L = 1$. Its maximum value (at $(1/2, 1/2)$) is $1/16$.

Let us begin by discussing an example of using a global (instead of piecewise) polynomial basis on Ω, which is possible when Ω is a square.

Example 2.7 (Uniformly heated square rod) Let us revisit Example 2.4 but now considering a square geometry, so that $\Omega = \{(x_1, x_2) \in \mathbb{R}^2, 0 < x_1 < L, 0 < x_2 < L\}$. The surface temperature is $g \in \mathbb{R}$ and the governing equation is $\Delta u = -f/k$ as before, where the heat source is $f \in \mathbb{R}$.

We want to select the space $\mathcal{W}_h$ as a subset of

$$\mathbb{P}_r(\Omega) = \{\text{polynomials of degree} \leq r \text{ in two variables in } \Omega\}\,. \tag{2.25}$$

If $r = 2$ this space consists of functions of the form

$$v(x_1, x_2) = c_1 + c_2 x_1 + c_3 x_2 + c_4 x_1^2 + c_5 x_1 x_2 + c_6 x_2^2\,,$$

if $r = 3$ the following terms are added

$$\ldots + c_7 x_1^3 + c_8 x_1^2 x_2 + c_9 x_1 x_2^2 + c_{10} x_2^3\,,$$

and so on. We will use this example to illustrate that the choice of a basis for $\mathcal{W}_h$ that can easily accommodate the constraints on $\mathcal{V}_h$ and $\mathcal{S}_h$ is not always trivial.

Both spaces $\mathcal{V}_h$ and $\mathcal{S}_h$ require functions to be constant along the boundary of the domain, that is, whenever x_1 or x_2 are either equal to 0 or equal to L. If, for example, we choose $r = 1$, then $p(x_1, x_2) = c_1 + c_2 x_1 + c_3 x_2$ for any $(c_1, c_2, c_3) \in \mathbb{R}^3$. The fact that p is constant at $x_2 = 0$ implies that

$\partial^i p/\partial x_1^i(x_1,0) = 0$ for all $i \in \mathbb{N}$ and any $x_1 \in [0,L]$. A similar argument can be made for the boundary conditions at $x_2 = L$, $x_1 = 0$ and $x_1 = L$. Hence,

$$\begin{aligned} 0 &= \frac{\partial p}{\partial x_1}(x_1,0) = c_2, \\ 0 &= \frac{\partial p}{\partial x_1}(x_1,L) = c_2, \\ 0 &= \frac{\partial p}{\partial x_2}(0,x_2) = c_3, \\ 0 &= \frac{\partial p}{\partial x_2}(L,x_2) = c_3, \end{aligned}$$

and in this case we only need to evaluate the first derivative. As a result, we conclude that functions in $\mathcal{V}_h$ or $\mathcal{S}_h$ need $c_2 = c_3 = 0$, so they can only be constant functions. Then, if $r = 1$, $\mathcal{V}_h = \{0 \cdot N_2\}$ and $\mathcal{S}_h = \{g \cdot N_2\}$, where $N_2(x_1,x_2) = 1$ constant for all $x \in \Omega$. These spaces contain a single function each, so they are poor choices for any approximation.

In a similar way, it can be verified that, if $r < 4$, $\mathcal{V}_h = \{0 \cdot N_2\}$ and $\mathcal{S}_h = \{g \cdot N_2\}$. For $r \geq 4$, due to the simplicity of the geometry, we have that the set of function in $\mathbb{P}_r(\Omega)$ that are constant on $\partial\Omega$ is (see the explanation after the example)

$$\mathcal{W}_h = \{v(x_1,x_2) = c_1 + \underbrace{x_1(L-x_1)x_2(L-x_2)}_{=0 \text{ on } \partial\Omega} p(x_1,x_2) \mid p \in \mathbb{P}_{r-4}, c_1 \in \mathbb{R}\},$$

and this will be our choice for $\mathcal{W}_h$. **Take** $r = 4$, which is the simplest case. Then $p(x_1,x_2)$ is a constant and $\mathcal{V}_h$ **has dimension 1**. Define the basis function

$$N_1(x_1,x_2) = x_1(L-x_1)x_2(L-x_2), \text{ so that } \nabla N_1 = \begin{pmatrix} (L-2x_1)x_2(L-x_2) \\ x_1(L-x_1)(L-2x_2) \end{pmatrix}.$$

Therefore, we can write

$$\mathcal{W}_h = \text{span}(\{N_1, N_2\}).$$

In particular, functions in $\mathcal{V}_h$ are zero on $\partial\Omega$, so

$$\mathcal{V}_h = \{v_1 N_1 \mid v_1 \in \mathbb{R}\}$$

and functions in $\mathcal{S}_h$ are equal to g on $\partial\Omega$, so

$$\mathcal{S}_h = \{v_1 N_1 + g N_2 \mid v_1 \in \mathbb{R}\}.$$

We then have $\eta_a = \{1\}$, $\eta_g = \{2\}$, we can choose $\overline{u}_h = g N_2$, and

$$u_h(x_1,x_2) = u_1 N_1(x_1,x_2) + g N_2(x_1,x_2) = g + u_1 N_1(x_1,x_2).$$

For u_h to satisfy the variational method with $a(\cdot,\cdot)$ given by (2.21c) and $\ell(\cdot)$ given by (2.21d) it must hold that

$$\int_\Omega k\nabla(g+u_1N_1)\cdot\nabla N_1\,dx_1dx_2=\int_\Omega f\,N_1\,dx_1dx_2\,.$$

Here we directly used that $u_2=g$, so we do not need to add an equation for the constrained index.

Noticing that $\nabla g=0$ and taking u_1 out of the integral by linearity, the final equation to compute $U=[u_1]$ is

$$KU=F$$

where the 1×1 stiffness matrix and load vector are

$$K=\int_\Omega k\nabla N_1\cdot\nabla N_1\,dx_1dx_2,\qquad F=\int_\Omega f\,N_1\,dx_1dx_2\,.$$

Performing the double integrals we obtain

$$K=\frac{kL^8}{45}\,,\quad F=\frac{fL^6}{36}\,,\quad\text{and thus}\quad u_1=\frac{5f}{4kL^2}\,.$$

This means that the solution of the variational method is

$$u_h(x_1,x_2)=g+\frac{5f}{4kL^2}\,N_1(x_1,x_2)=g+\frac{5f}{4kL^2}\,x_1(L-x_1)x_2(L-x_2)\,.$$

The maximum temperature takes place at the center (if $f>0$), with value $g+5fL^2/(64k)$. The temperature contours are shown in Fig. 2.6, where we compare the numerical solution (obtained by solving an equation with just one unknown!) with the exact solution. They are qualitatively very similar. The maximum difference is, in fact, less than 6%.

It is important to remark that the numerical solution u_h is **not** an exact solution of the differential equation. To check this, simply compute

$$\Delta u_h(x_1,x_2)=-\frac{5f}{2kL^2}\,[x_1(L-x_1)+x_2(L-x_2)]$$

and compare to the exact equation $\Delta u=-f/k$. The **residual of the differential equation**, evaluated on the numerical solution, is

$$r_h=\Delta u_h(x_1,x_2)+\frac{f}{k}=\frac{f}{k}\left(1-\frac{5}{2}\,\frac{x_1(L-x_1)+x_2(L-x_2)}{L^2}\right).$$

Its average value over the domain is $f/(6k)$. The residual is maximum at the vertices, where its value is f/k. At the center the value is $-1/4$. The residual is plotted in Fig. 2.7. It must not be mistaken for the actual **error of the approximate solution** $e_h=u-u_h$, which in this case we can compute because the exact solution is known and is also shown in the same figure.

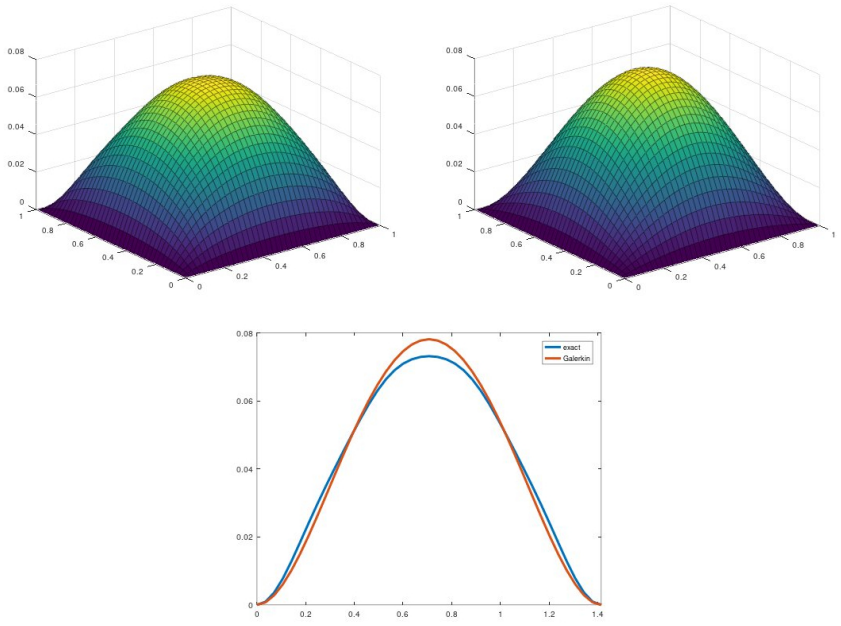


Figure 2.6 Exact solution, variational method solution, and comparison along the diagonal.

Functions in $\mathbb{P}_r(\Omega)$, $r \geq 4$ that are constant on $\partial\Omega$, where $\Omega = [0,L] \times [0,L]$

To obtain the result, we will repeatedly use the following observation. Let $q \in \mathbb{P}_r(\Omega)$, and subtract a constant c_1 so that $q_r(x_1,x_2) = q(x_1,x_2) - c_1$ is equal to zero on $\partial\Omega$. Let $h(x_1,x_2) = h_0 + h_1 x_1 + h_2 x_2$ for $h_0, h_1, h_2 \in \mathbb{R}$ such that either h_1 or h_2 are not zero, and such that if $h(\bar{x}_1,\bar{x}_2) = 0$ then $q(\bar{x}_1,\bar{x}_2) = 0$. Then,

$$q_r(x_1,x_2) = h(x_1,x_2)\, q_{r-1}(x_1,x_2), \tag{2.26}$$

where $q_{r-1}(x_1,x_2)$ is a polynomial of degree $r-1$. To see this, without loss of generality assume that $h_1 \neq 0$, and let $z = h(x_1,x_2)$, so that $x_1 = g(z,x_2) = (z - h_2 x_2 - h_0)/h_1$. Consider then the polynomial $\hat{q}_r(z,x_2) = q_r(g(z,x_2),x_2)$, which satisfies that $\hat{q}_r(0,x_2) = 0$ for any x_2. Then, it admits a factorization of the form

$$\hat{q}_r(z,x_2) = z\hat{q}_{r-1}(z,x_2)$$

for a polynomial $\hat{q}_{r-1}$ of degree $r-1$. Defining $q_{r-1}(x_1,x_2) = \hat{q}_{r-1}(h(x_1,x_2),x_2)$, we arrive to (2.26).

We can apply this to our case, by sequentially considering $h(x_1,x_2)$ equal to $x_1, L - x_1, x_2$ and $L - x_2$. It then follows that

$$q_r(x_1,x_2) = x_1(L-x_1)x_2(L-x_2)\, q_{r-4}(x_1,x_2), \tag{2.27}$$

where q_{r-4} is a polynomial of degree $r-4$.

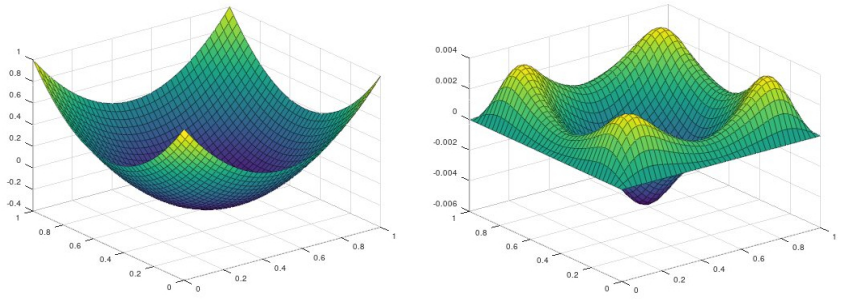


Figure 2.7 The residual function $r_h(x_1, x_2)$ (left) and the error function $e_h(x_1, x_2)$ (right).

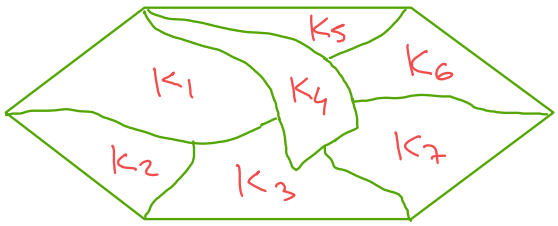


Figure 2.8 Example of a mesh with $n_{\text{el}} = 7$

2.4 Finite Element Spaces in Two Dimensions

Over the years many finite element spaces have been introduced with ever-increasing sophistication for different specific applications. For diffusion problems the classical and still most popular ones consist of **piecewise polynomial functions that are continuous in** Ω. These are the ones that we discuss next.

2.4.1 The Simplest C^0 Finite Element Space in Two Dimensions

How to define and build a space of piecewise polynomials that only consist of continuous functions? Consider the domain subdivided into element subdomains, over which the functions of the space need to be polynomials of degree r in two variables. If r is zero, the function is continuous if and only if its value is the same in all element subdomains, which makes the *piecewise* constant functions to be *globally* constant. So, $r = 0$ does not produce a space of continuous functions that can be used to approximate anything.

But what about piecewise *linear* polynomials ($r = 1$)? Or polynomials of higher degree? To consider the simplest case, lets answer this question for $r = 1$. The answer will introduce us to the **continuous** P_1 **finite element space**, also known simply as P_1 space. To introduce this space, let's first define what a mesh is.

In the following, an **element domain** is a set in $\mathbb{R}^2$ (or $\mathbb{R}^3$ in 3D) that can be obtained by a bijective deformation of a disc (or a sphere in 3D).

Definition 2.1 (Mesh)**.** *A* **mesh** $\mathcal{T} = \{K_1, \ldots, K_{n_{el}}\}$ *for a domain* Ω *is a collection of a finite number* n_{el} *of element domains* $K_1, \ldots, K_{n_{el}}$ *such that* $\mathring{K}_i \cap \mathring{K}_j = \emptyset$ *and* $\Omega = \cup_{i=1}^{N_{el}} K_i$.

The symbol $\mathring{\omega}$ for $\omega \subset \mathbb{R}^2$ denotes the *interior* of ω. A point $x \in \mathring{\omega}$ if we can find a disc D centered at x such that $D \subset \omega$, that is, D is completely included in ω, see Fig. 2.9.

Figure 2.8 shows an example of a mesh. When all elements in a mesh are triangles (or tetrahedra in 3D), the mesh is also called a **triangulation**. **Vertices and edges** of a triangulation are any of the vertices and/or edges of the triangles in the triangulation.

Definition 2.2 (Continuous P_1 finite element space)**.** *Given a mesh* $\mathcal{T}$ *for a domain* Ω*, the* **continuous** P_1 **finite element space** *over* $\mathcal{T}$ *is the space* $\mathcal{W}_h$ *of all functions that belong to* $C^0(\Omega)$ *and are a polynomial of degree* $r = 1$ *in each of the element domains.*

A polynomial $p(x_1, x_2)$ of degree less or equal than 1 in two dimensions has the form

$$p(x_1, x_2) = c_1 + c_2 x_1 + c_3 x_2$$

Therefore, if we know the value of p at three different non-colinear points in the plane, we can uniquely identify the coefficients c_1, c_2 and c_3.

Given any partition of Ω it may be very hard to exhibit a basis of the P_1 space, but there are some special meshes, called **conforming triangulations** that make it very simple.

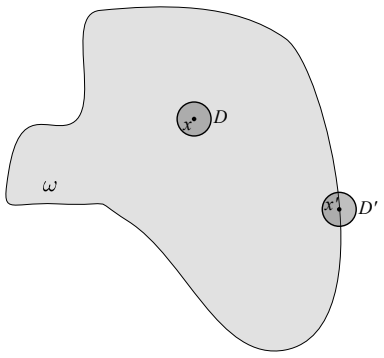


Figure 2.9 The point x belongs to the interior of ω, while x' does not.

Definition 2.3 (Conforming triangulation)**.** *A* **conforming triangulation** *of a polygonal domain* Ω *is a mesh for* Ω *such that the intersection of any two triangles* K *and* K' *is either (a) empty, or (b) a whole edge of* **both** K *and* K'*, or (c) a vertex of* **both** K *and* K'.

Figure 2.10 shows several meshes for a polygonal domain. The one labeled B is a conforming triangulation. The ones labeled A and C are not conforming.

Conforming triangulations are remarkable. Let $\mathcal{T}$ be a triangulation and let us number the vertices of the triangulation from 1 to n_V. Since every triangle has three vertices, by specifying the values $f_1, f_2, \ldots, f_{n_V}$ at all vertices of $\mathcal{T}$ one defines a *unique* $\mathbb{P}_1$-polynomial function f^K in each triangle K of $\mathcal{T}$. From these functions $\{f^K\}$, each defined in one triangle of $\mathcal{T}$, we can build a function f over Ω as

$$f(x) = f^K(x)\,, \quad \text{if } x \in K\,. \tag{2.28}$$

In other words, f takes at a point x the value corresponding to $f^K(x)$, if the point belongs to triangle K. All points belong to at least one triangle, so that $f(x)$ always exists. However, at some edges of a general triangulation the function may be multi-valued and thus f **may not be a continuous function**. This is easily understood by considering, in Fig. 2.10 (case A), the function f that is zero at all vertices except for the one and only interior vertex, at which the value is 1. This function, drawn in Fig. 2.11, is discontinuous along the edge over which the interior vertex is "hanging".

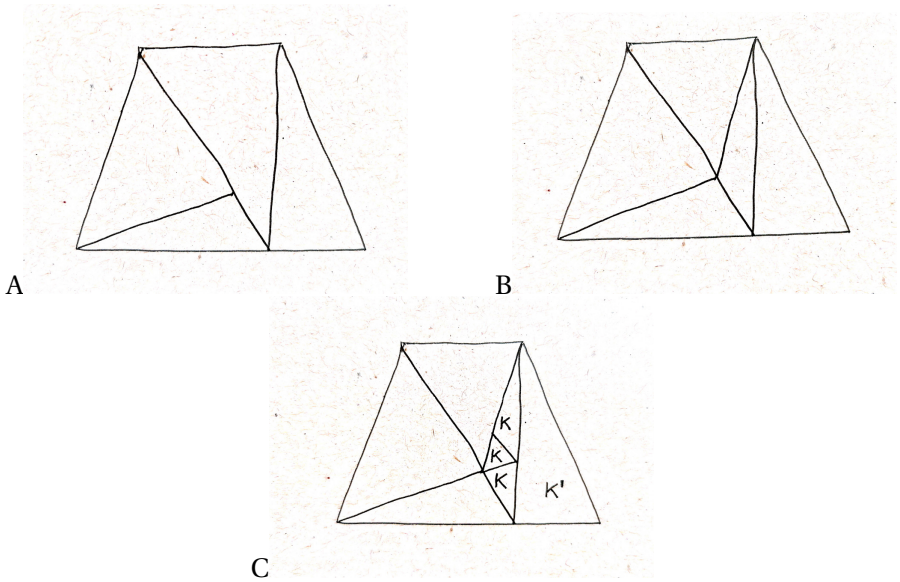


Figure 2.10 Examples of triangulations of a polygonal domain. B is conforming, A and C are not. The elements K in triangulation C that fail the definition of conforming triangulation when considered vis-à-vis K' are indicated.

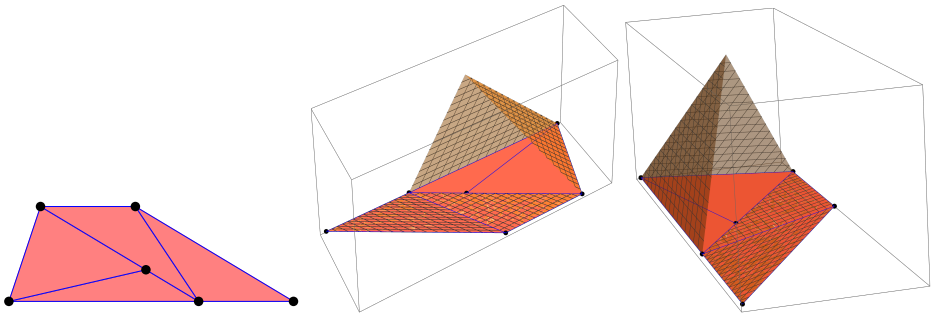


Figure 2.11 The non-conforming mesh in Fig. 2.10(Case A) is shown on the left, and two views of a piecewise linear polynomial function over it are displayed at the center and on the right. The function is obtained by setting the value of the interior vertex to 1, and to 0 at all the remaining vertices. Because the triangulation is not conforming, the resulting function is discontinuous along one of the interior edges.

Definition 2.4 (Hanging vertex)**.** *A* **hanging vertex** *in a triangulation $\mathcal{T}$ is a vertex v for which there exists a triangle K in $\mathcal{T}$ satisfying*

$$v \in K, \qquad \textit{and} \qquad v \textit{ is not a vertex of } K\,.$$

An alternative definition of **conforming triangulations** is **triangulations with no hanging vertices**.

The remarkable fact about conforming triangulations is the following

Theorem 2.2. *Let $\mathcal{T}$ be a conforming triangulation with n_V vertices. Given arbitrary scalar values $f_1, f_2, \ldots, f_{n_V}$ at the vertices, then*

- *the piecewise $\mathbb{P}_1$ function f defined by (2.28) is* **always continuous**, *and*
- *every function in the space $\mathcal{W}_h$ of all continuous piecewise $\mathbb{P}_1$ functions corresponds to a specific set of vertex values $(f_1, f_2, \ldots, f_{n_V}) \in \mathbb{R}^{n_V}$.*

As a consequence, the set of functions $\{N_a,\ a = 1, \ldots, n_V\}$, each one constructed as in (2.28), in which N_a takes the value 1 at vertex a and the value zero at all other vertices, **is a basis of** $\mathcal{W}_h$.

If the position of vertex i is $x_i = (x_{i1}, x_{i2})$, then the **Krönecker-delta property** holds, i.e.,

$$N_j(x_i) = \delta_{ij} \tag{2.29}$$

with $\delta_{ij} = 1$ if $i = j$ and $= 0$ otherwise.

Any function w_h in $\mathcal{W}_h$ is a linear combination of the basis functions. The coefficients of the linear combination, or **degrees of freedom**, are the values at the vertices, which by such property are called **nodes of the P_1 finite element space**. We thus have $n_V = m$ and

$$w_h(x) = w_1 N_1(x) + w_2 N_2(x) + \ldots + w_m N_m(x)\,, \tag{2.30}$$

where

$$w_1 = w_h(x_1), \quad w_2 = w_h(x_2), \qquad \ldots \qquad w_m = w_h(x_m)\,. \tag{2.31}$$

Once more we observe the linear correspondence between the **column array** of nodal values $\mathsf{W} = (w_1, w_2, \ldots, w_m)^T$ and the **function** $w_h(x)$.

How is a triangulation handled within a code? The typical way in which triangulations are handled is by means of two basic arrays:

- The **vertex coordinates array**, denoted here by X. It is a matrix of n_V columns, such that each column is the coordinate vector of a vertex.
- The **list of vertices**, denoted here by LV. It is a matrix of n_T columns and 3 rows. Each column contains the three numbers identifying the three vertices of the corresponding triangle. This is also called the **connectivity array**.

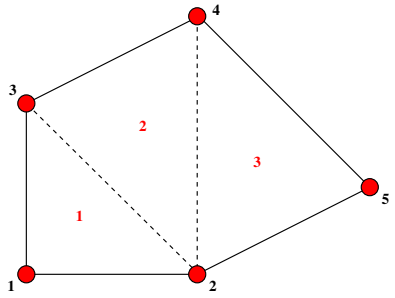


Figure 2.12 A simple conforming triangulation

Example 2.8 (A simple conforming triangulation) The triangulation of Fig. 2.12 has $n_V = 5$ vertices and $n_T = 3$ triangles and is conforming, as can easily be checked. With the specified numbering of triangles and vertices, the corresponding arrays are

$$\mathsf{X} = \begin{pmatrix} 4 & 8 & 4 & 8 & 12 \\ 2 & 2 & 6 & 8 & 4 \end{pmatrix} \tag{2.32}$$

$$\mathsf{LV} = \begin{pmatrix} 1 & 3 & 4 \\ 2 & 2 & 2 \\ 3 & 4 & 5 \end{pmatrix} \tag{2.33}$$

The basis functions $N_1, \ldots, N_5$ of the P_1 space corresponding to this triangulation can be seen in Fig. 2.13.

2.4.1.1 Barycentric or Area Coordinates

The **geometry** of a P_1 triangle K is determined by the positions its vertices $\mathbf{X}^1$, $\mathbf{X}^2$ and $\mathbf{X}^3$. It is the **only** triangle that has such vertices. It is also the **convex hull** of $\{\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3\}$, defined as the convex linear combinations (CLC) of the vertex positions:

$$K = \mathscr{C}(\{\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3\}) = \left\{ \sum_{j=1}^{d+1} \lambda_j \mathbf{X}^j \;\middle|\; \underbrace{0 \le \lambda_j \le 1 \;\forall j, \text{ and } \sum_{j=1}^{d+1} \lambda_j = 1}_{\hat{K}} \right\} \tag{2.34}$$

The set $\hat{K}$ defines a triangle, called **reference triangle**, illustrated in Fig. 2.14.
Remark: This definition of the geometry, in fact, works equally well in 2D ($d = 2$) and 3D ($d = 3$, in which case the triangle turns into a tetrahedron). It is independent of d. For each $\mathbf{x} \in K$ there exists a unique triplet $(\lambda_1, \lambda_2, \lambda_3) \in \hat{K}$ such that $\mathbf{x} = \sum_{j=1}^{3} \lambda_j \mathbf{X}^j$, thus

$$\mathbf{x} \in K \iff (\lambda_1, \lambda_2, \lambda_3) \in \hat{K}$$

is one-to-one, and thus a **reparameterization** (change of coordinates) of K. The parameters λ_i are called **area coordinates** of K (also **barycentric coordinates**).

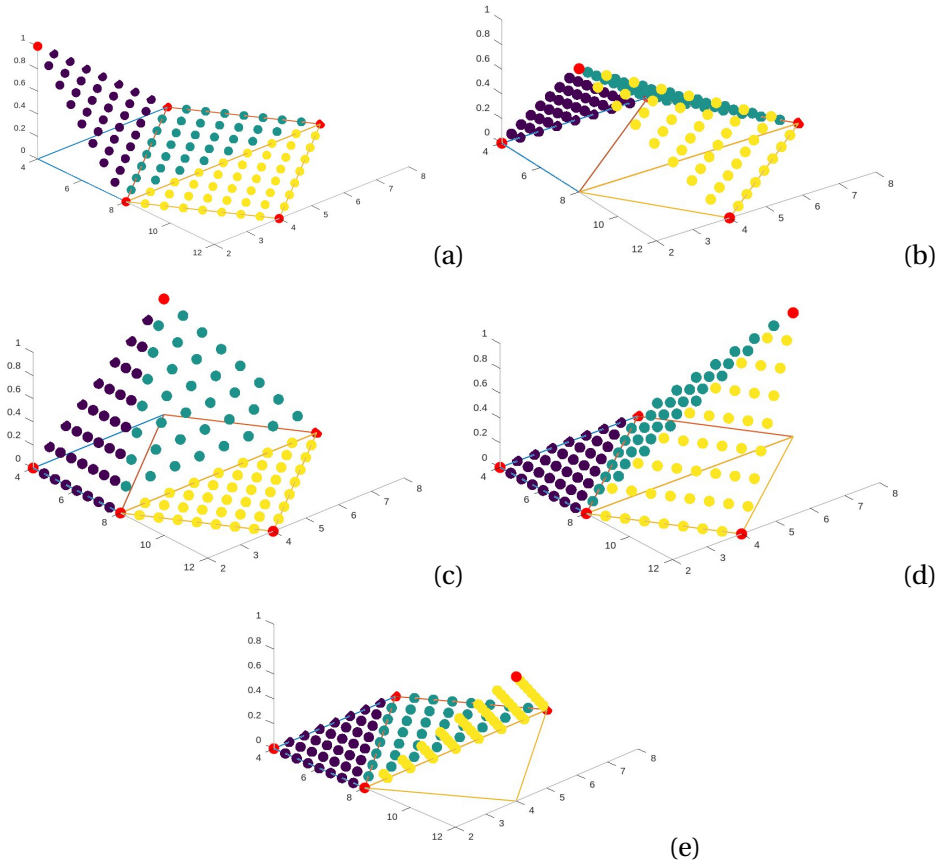


Figure 2.13 Basis functions of the triangulation of Example 2.8. (a) N_1, (b) N_2, ..., (e) N_5. The elevation corresponds to the value of the function. The nodal values are shown as red dots. Sample points belonging to triangle 1, 2 or 3 are shown in purple, green or yellow, respectively.

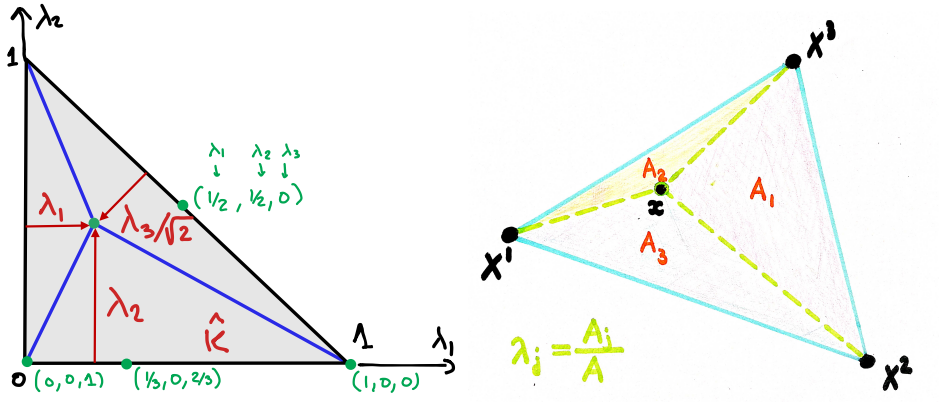


Figure 2.14 Left: Reference triangle $\hat{K}$ and domain of the barycentric coordinates. Right: Interpretation of barycentric coordinates in a general triangle as the fraction of the area of the triangle formed by a point x in the triangle and two of the vertices.

Barycentric coordinates have the following properties:

a) $\mathbf{x} = \mathbf{X}^j \iff \lambda_j = 1, \lambda_{k\neq j} = 0$. This is analogous to the Krönecker delta property.

b) $\mathbf{x}$ belongs to edge $\mathbf{X}^i\mathbf{X}^j$ iff $\lambda_i + \lambda_j = 1$, and as a result, $\lambda_{k\notin\{i,j\}} = 0$. In other words, λ_j is zero along the edge opposite to vertex $\mathbf{X}^j$.

c) $\mathbf{x}$ belongs to the (relative) interior of K iff all λ_j's are different from 0 and 1.

d) The barycentric coordinates satisfy that

$$\lambda_i = \frac{A_i}{A},$$

where A is the area of the triangle K and A_i is the area of the triangle formed by $\mathbf{x}$ and the two vertices $\mathbf{X}^j$ with $j \neq i$, see Fig. 2.14.

e) The inverse mapping to $(\lambda_1, \lambda_2, \lambda_3) \mapsto \mathbf{x} = \sum_{j=1}^3 \lambda_j \mathbf{X}^j$ is given by (in 2D, with $\mathbf{x} = (x_1, x_2)^T$)

$$\lambda_1(x_1, x_2) = \frac{1}{2A}\left[-(X_2^3 - X_2^2)(x_1 - X_1^2) + (X_1^3 - X_1^2)(x_2 - X_2^2)\right] \quad (2.35a)$$

$$\lambda_2(x_1, x_2) = \frac{1}{2A}\left[-(X_2^1 - X_2^3)(x_1 - X_1^3) + (X_1^1 - X_1^3)(x_2 - X_2^3)\right] \quad (2.35b)$$

$$\lambda_3(x_1, x_2) = \frac{1}{2A}\left[-(X_2^2 - X_2^1)(x_1 - X_1^1) + (X_1^2 - X_1^1)(x_2 - X_2^1)\right] \quad (2.35c)$$

where $2A$ is twice the area of K,

$$2A = (X_1^2 - X_1^1)(X_2^3 - X_2^1) - (X_2^2 - X_2^1)(X_1^3 - X_1^1)\,.$$

It is a general convention that the *vertices are ordered either clockwise or counter-clockwise*. We are adopting the latter. Otherwise, A would be negative the area of K, but the other formulae would remain true.

f) The **barycenter B** of K corresponds to

$$\mathbf{B} = \frac{\mathbf{X}^1 + \mathbf{X}^2 + \mathbf{X}^3}{3} \leftrightarrow (\lambda_1, \lambda_2, \lambda_3) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$$

g) The **edges midpoints** correspond to $(\lambda_1, \lambda_2, \lambda_3)$ equal to $(0, 1/2, 1/2)$, $(1/2, 0, 1/2)$ and $(1/2, 1/2, 0)$. Notice that the midpoints have been numbered according to the opposite vertex.

Inversion of the barycentric coordinate map.

We obtain formulas (2.35) next. For a triangle, the map $(\lambda_1, \lambda_2, \lambda_3) \mapsto (x_1, x_2)$ is defined by the following equations:

$$\begin{aligned} x_1 &= \lambda_1 X_1^1 + \lambda_2 X_1^2 + \lambda_3 X_1^3 \\ x_2 &= \lambda_1 X_2^1 + \lambda_2 X_2^2 + \lambda_3 X_2^3 \\ 1 &= \lambda_1 \quad + \lambda_2 \quad + \lambda_3. \end{aligned}$$

If we know (x_1, x_2), we can solve for $(\lambda_1, \lambda_2, \lambda_3)$ by solving this system of 3 equations with 3 unknowns. The determinant of this system is

$$\begin{vmatrix} X_1^1 & X_1^2 & X_1^3 \\ X_2^1 & X_2^2 & X_2^3 \\ 1 & 1 & 1 \end{vmatrix}.$$

This determinant represents the twice the (signed) volume of a tetrahedron formed by the origin and the triangle placed on the horizontal $x_1 - x_2$ plane defined by $x_3 = 1$. Since it has height equal to 1, its volume is non-zero if and only the area of the triangle is not zero. Therefore, this system is invertible, and the solution is given by (2.35).

2.4.1.2 The P_1-Element and the Local-to-Global Map

The barycentric coordinates are not just another set of coordinates (instead of $x_1 - x_2$) that one could choose to parameterize the points of a triangle, for **triangular finite elements**,

$$N_1^e = \lambda_1, \quad N_2^e = \lambda_2 \quad \text{and} \quad N_3^e = \lambda_3, \tag{2.36}$$

given by (2.35a)-(2.35c) above, are **the local basis of the P_1-finite element space** restricted to element e. In fact, they are **three polynomials of degree 1 and linearly independent**, and thus a basis of $\mathbb{P}_1$. Since they satisfy the Krönecker-delta property at the vertices, the **vertices are the nodes of this space**.

We can then define a P_1-element as $e = (\Omega_e, \{N_1^e, N_2^e, N_3^e\})$, where Ω_e is a triangle.

The gradient of the basis functions can be obtained by differentiation of (2.35a)-(2.35c) with respect to x_1 and x_2, which gives

$$\nabla N_1^e = \frac{1}{2A} \begin{pmatrix} X_2^2 - X_2^3 \\ X_1^3 - X_1^2 \end{pmatrix}, \tag{2.37}$$

$$\nabla N_2^e = \frac{1}{2A} \begin{pmatrix} X_2^3 - X_2^1 \\ X_1^1 - X_1^3 \end{pmatrix}, \tag{2.38}$$

$$\nabla N_3^e = \frac{1}{2A} \begin{pmatrix} X_2^1 - X_2^2 \\ X_1^2 - X_1^1 \end{pmatrix}. \tag{2.39}$$

The second derivatives are of course zero all over the element.

Now, **notice what happens if we take the local-to-global array equal to the list-of-vertices array**,

$$\mathsf{LG} = \mathsf{LV}. \tag{2.40}$$

This means that we consider the P_1 element with **the vertices as nodes and the triangles as element domains**.

Following exactly the same methodology that was developed for the 1D case, the global basis functions are defined as

$$N_A(x_1, x_2) = \mathop{\sum_{\{(a,e)|\mathsf{LG}(a,e)=A\}}}^{\circ} N_a^e(x_1, x_2). \tag{2.41}$$

Remember, the summation is only performed when $x = (x_1, x_2)$ is **interior** to some triangle in the mesh. The value of N_A at the mesh vertices and edges is not equal to the sum above, but as the continuous extension (if it exists) from the element interiors. This is formalized in the following definition.

Definition 2.5 (Broken Sum)**.** *Let $\mathcal{T} = \{K_1, \ldots, K_{n_{el}}\}$ be a mesh for a domain $\Omega \subset \mathbb{R}^d$. Let $\mathcal{W}_h$ be a space of scalar-valued functions over Ω such that if $f_h \in \mathcal{W}_h$, then f_h is continuous in the interior of each element domain. The broken sum $\mathring{+}: \mathcal{W}_h \times \mathcal{W}_h \to \mathcal{W}_h$ is defined for $f_h, g_h \in \mathcal{W}_h$ by*

$$(f_h \mathring{+} g_h)(x) = f_h(x) + g_h(x), \qquad x \in \bigcup_{i=1}^{n_{el}} \mathring{K}_i \tag{2.42a}$$

and

$$(f_h \mathring{+} g_h)(x) = \lim_{y \to x} f_h(y) + g_h(y), \qquad \textit{otherwise.} \tag{2.42b}$$

Let us see how this works in the triangulation of Figure 2.12.

Example 2.9 (Using the list of vertices as local-to-global map) Going back to the triangulation in Example 2.8, with LG equal to LV given in (2.33), i.e.,

$$\mathsf{LG} = \mathsf{LV} = \begin{pmatrix} 1 & 3 & 4 \\ 2 & 2 & 2 \\ 3 & 4 & 5 \end{pmatrix}, \tag{2.43}$$

the explicit expressions for the global basis functions (corresponding to (2.41)) are

$$N_1 = N_1^1 \tag{2.44}$$
$$N_2 = N_2^1 \mathring{+} N_2^2 \mathring{+} N_2^3 \tag{2.45}$$
$$N_3 = N_3^1 \mathring{+} N_1^2 \tag{2.46}$$
$$N_4 = N_3^2 \mathring{+} N_1^3 \tag{2.47}$$
$$N_5 = N_3^3 \tag{2.48}$$

Take for example N_3, which is depicted in Fig. 2.15. We know that N_3 is different from zero just in elements 1 and 2 because in LG the number 3 only appears in columns 1 and 2. Inside element $e = 1$ the function N_3 coincides with the N_3^1, depicted in purple in the figure, because 3 appears in row 3 of column $e = 1$ of LG. Similarly, in element $e = 2$ the function N_3 coincides with N_1^2, depicted in green in the figure, because 3 appears in row 1 of column $e = 2$ of LG. In column $e = 3$ of LG the number 3 does not appear. This means that the function N_3 is identically zero in element $e = 3$, as shown in yellow in Fig. 2.15. By the magic of conforming triangulations, the three pieces fit together in such a way that the resulting function N_3 is a continuous function. **In fact, the functions N_1-N_5 defined by (2.44)-(2.44) are exactly the same as those defined in Theorem 2.2 and depicted in Figure 2.13.**

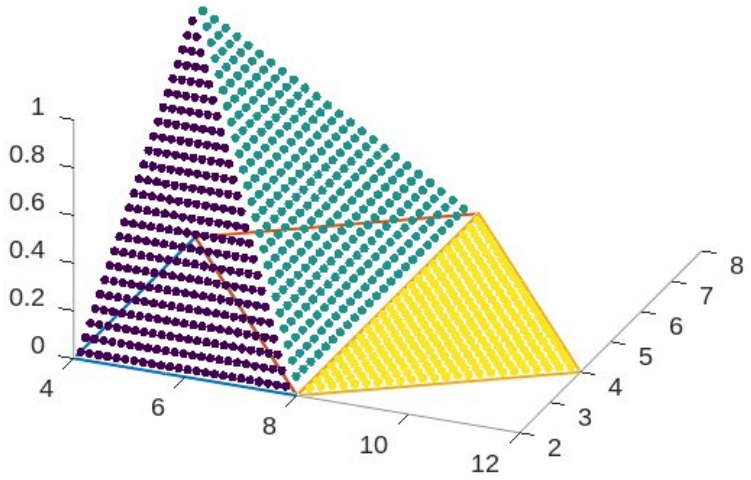


Figure 2.15 The basis function N_3 sampled only at interior points. Sample points belonging to triangle 1, 2 or 3 are shown in purple, green or yellow, respectively.

The fundamental message is that, if the arrays X and LV of a conforming triangulation of a domain Ω are available, then the global basis functions obtained from (2.41) by taking LG = LV generate the space $\mathcal{W}_h$ of P_1 continuous finite elements.

2.4.2 Elemental Computations with the Simplest C^0 Space

In what follows, we proceed and use the basis functions just developed to solve a 2D diffusion problem in any polygonal domain by a variational numerical method.

2.4.2.1 The Element Stiffness Matrix

Remember that our goal is to solve the diffusion Problem 2.2 based on variational equation (2.21) with $\mathcal{W}_h$ a finite element space over a conforming triangulation for Ω, so as to obtain the only function $u_h \in \mathcal{S}_h$ that satisfies

$$\int_\Omega (K\nabla u_h)\cdot\nabla v_h \, d\Omega = \int_\Omega f\, v_h \, d\Omega + \int_{\partial\Omega_N} H\, v_h \, d\Gamma \tag{2.49}$$

for all $v_h \in \mathcal{V}_h$.

The element stiffness matrix of the diffusion problem for the P_1 triangular element is as follows. Let the indices a and b run from 1 to 3, the number of nodes per element. The three local basis functions at element e are barycentric coordinates, as introduced in (2.36). As a consequence, the element matrix results from

$$\mathsf{K}^e_{ab} = \int_{\Omega_e} (K\nabla N^e_b)\cdot\nabla N^e_a \, d\Omega \,. \tag{2.50}$$

Let us assume for simplicity that K is isotropic, so that $K(x) = k(x)\mathbb{I}$. Further, notice that ∇N_b^e is constant over Ω_e for any b, so that

$$\mathsf{K}_{ab}^e = \left(\int_{\Omega_e} k(x)\, d\Omega\right) \nabla N_b^e \cdot \nabla N_a^e = k_e\, A_e\, \nabla N_b^e \cdot \nabla N_a^e\,.$$

Notice that k_e is the average diffusion coefficient in the element, since A_e is the element's area.

Consider, for each element e, the array dN which contains the partial derivatives of the basis functions and is defined by

$$\mathsf{dN}_{ib} = \frac{\partial N_b^e}{\partial x_i}\,,$$

that is, one column per basis function, with $i = 1,2$. Then K^e is given by

$$\mathsf{K}^e = k_e\, A_e\, \mathsf{dN}^T\, \mathsf{dN}\,.$$

☞ The matrix dN has the form

$$\mathsf{dN} = \begin{bmatrix} \frac{\partial N_1^e}{\partial x_1} & \frac{\partial N_2^e}{\partial x_1} & \frac{\partial N_3^e}{\partial x_1} \\ \frac{\partial N_1^e}{\partial x_2} & \frac{\partial N_2^e}{\partial x_2} & \frac{\partial N_3^e}{\partial x_2} \end{bmatrix}.$$

In Octave/MATLAB code, given the array of nodal coordinates of element e,

$$\mathsf{xe} = \begin{pmatrix} X_1^1 & X_1^2 & X_1^3 \\ X_2^1 & X_2^2 & X_2^3 \end{pmatrix}$$

in which each column corresponds to one of the three nodes of the element, and given k_e, the computation of K^e is as simple as

```
dN=[xe(2,2)-xe(2,3),xe(2,3)-xe(2,1),xe(2,1)-xe(2,2);...
      xe(1,3)-xe(1,2),xe(1,1)-xe(1,3),xe(1,2)-xe(1,1)];
Ae2=dN(2,3)*dN(1,2)-dN(1,3)*dN(2,2);
dN=dN/Ae2;
Ke=Ae2/2*ke*dN'*dN;
```

2.4.2.2 The Element Load Vector

Let us for now assume that either Dirichlet conditions are imposed all over the boundary or the Neumann datum H is zero. Then the element load vector is

$$\mathsf{F}_a^e = \int_{\Omega_e} f\, N_a^e\, d\Omega\,.$$

The result of this integral depends of course on the function $f(x)$. Let us assume that $f = f_e$ is constant over the element, then $f N_a^e$ is a polynomial of degree 1 in x_1 and x_2 which has as integral

$$\mathsf{F}_a^e = f_e \int_{\Omega_e} N_a^e\, d\Omega = \frac{f_e A_e}{3}, \qquad \forall a = 1,2,3. \tag{2.51}$$

The computation of the element load vector is thus immediate

```
Fe=Ae2*fe*ones(3,1)/6;
```

Another possibility is to assume that f is affine within the element, in which case the three nodal values need to be provided. Whatever the values are, it is clear that the product $f\,N_a^e$ is a polynomial of degree 2 in the variables x_1 and x_2.

To compute F^e we can use the following well-known integration rule: *The integral of a quadratic polynomial over a triangle is equal to $\frac{1}{3}$ times the triangle's area times the sum of the values of the polynomial at the three edge midpoints.*

Let $f_e = (f_{e,1}, f_{e,2}, f_{e,3})$ be an array containing the three nodal values of f_e. Then the integration rule above states that

$$\begin{aligned}\mathsf{F}_a^e &= \int_{\Omega_e} f\,N_a^e\,d\Omega \\ &= \frac{A_e}{3}\left(\frac{f_{e,1}+f_{e,2}}{2}\,\frac{\delta_{a1}+\delta_{a2}}{2} + \frac{f_{e,2}+f_{e,3}}{2}\,\frac{\delta_{a2}+\delta_{a3}}{2} + \frac{f_{e,3}+f_{e,1}}{2}\,\frac{\delta_{a3}+\delta_{a1}}{2}\right)\end{aligned}$$

Particularizing for $a = 1,2,3$ we obtain the element load vector

$$\mathsf{F}^e = \frac{A_e}{12}\begin{pmatrix} 2f_{e,1}+f_{e,2}+f_{e,3} \\ f_{e,1}+2f_{e,2}+f_{e,3} \\ f_{e,1}+f_{e,2}+2f_{e,3}\end{pmatrix} \tag{2.52}$$

which would lead to the code (notice that `fe` is stored as a row array)

```
auxmat=[2,1,1;1,2,1;1,1,2];
Fe=(Ae2/24)*auxmat*fe';
```

2.4.3 Solving Problems with Dirichlet Boundaries

Let us work out the procedure and code that solves the diffusion problem with P_1 finite elements, assuming that we only have Dirichlet boundary conditions ($\partial\Omega_D = \partial\Omega$). We assume, as before, that a mesh is provided by means of a **list of coordinates** X and a **list of vertices** LV. This enables us to construct $\mathcal{W}_h$ as the continuous P_1 finite element space over it. The dimension m of this space is equal to the number of vertices n_V, so $m = n_V$.

The trial and test spaces are then constructed as

$$\begin{aligned}\tilde{\mathcal{S}}_h &= \{w_h \in \mathcal{W}_h \mid w_h(x) = g(x) \quad \forall\, x \in \partial\Omega_D\} \\ \tilde{\mathcal{V}}_h &= \{w_h \in \mathcal{W}_h \mid w_h(x) = 0 \quad \forall\, x \in \partial\Omega_D\}.\end{aligned}$$

As in §1.3.1.1, the trial space is set to be the subset of functions in $\mathcal{W}_h$ that satisfy the essential boundary conditions, and the test space is the direction of the trial space. There is a caveat here though: functions in $\mathcal{W}_h$ are affine on each edge in $\partial\Omega$, but g may not be! As a result, the space $\tilde{\mathcal{S}}_h$ may have *no* functions in it ...

Fortunately, it is not necessary to satisfy essential boundary conditions exactly for convergence, but it is enough to approximate them. A suitable way to do it in this case is by requiring w_h to be equal to g at each vertex (and hence node for P_1 element) on $\partial\Omega$, namely,

$$\mathcal{S}_h = \{w_h \in \mathcal{W}_h \mid w^a = g(\mathbf{X}^a) \text{ for each vertex } \mathbf{X}^a \in \partial\Omega_D\} \tag{2.53a}$$

$$\mathcal{V}_h = \{w_h \in \mathcal{W}_h \mid w^a = 0 \text{ for each vertex } \mathbf{X}^a \in \partial\Omega_D\} \tag{2.53b}$$

Once more, we defined the test space $\mathcal{V}_h$ as the direction of $\mathcal{S}_h$. Notice, that $\mathcal{V}_h = \tilde{\mathcal{V}}_h$. A basis for $\mathcal{V}_h$ is obtained by considering all basis functions in $\mathcal{W}_h$ that are 0 on $\partial\Omega_D$. To see this, notice that

$$w_h(x) = 0 \text{ for } all\ x \in \partial\Omega_D \iff w^a = 0 \text{ whenever } \mathbf{X}^a \in \partial\Omega_D.$$

Hence, if vertex $\mathbf{X}^a \in \partial\Omega_D$, then $N_a \notin \mathcal{V}_h$. Instead, if $\mathbf{X}^a \notin \partial\Omega_D$, then $N_a \in \mathcal{V}_h$. Once again, we should stop and reflect on the ease by which it is possible to identify a subset of the finite element basis for $\mathcal{W}_h$ to serve as basis for $\mathcal{V}_h$. This is not trivial or even possible with many other bases, as Example 2.7 illustrates.

So, the set of active and constrained indices is

$$\eta_a = \{a \in \{1,\dots,n_V\} \mid \mathbf{X}^a \notin \partial\Omega_D\}$$
$$\eta_g = \{a \in \{1,\dots,n_V\} \mid \mathbf{X}^a \in \partial\Omega_D\}.$$

This definition of $\mathcal{S}_h$ automatically identifies a function $\overline{u}_h \in \mathcal{S}_h$ as

$$\overline{u}_h = \sum_{a\in\eta_g} g(\mathbf{X}^a) N_a.$$

However, in this case the equations that constrained indices should satisfy are already included in the definition of $\mathcal{S}_h$, namely,

$$u^a = \overline{u}^a = g(\mathbf{X}^a) \qquad a \in \eta_g. \tag{2.54}$$

We can assume that a **list of boundary nodes**, those with indices in η_g, is provided by the way the mesh is constructed. The **list of boundary values**, GG, should be constructed from (2.54).

Finally, at this stage we can approximate k and f as piecewise constant over each triangle, so that we count with values k_e and f_e in each triangle.

2.4.3.1 Element Contributions

We can build a function that computes the element stiffness matrix and element load vector, for example:

```
function [Ke, Fe]=elementKandF(xe,ke,fe)
  dN=[xe(2,2)-xe(2,3),xe(2,3)-xe(2,1),xe(2,1)-xe(2,2);...
      xe(1,3)-xe(1,2),xe(1,1)-xe(1,3),xe(1,2)-xe(1,1)];
  Ae2=dN(2,3)*dN(1,2)-dN(1,3)*dN(2,2);
  dN=dN/Ae2;
  Ke=Ae2/2*ke*dN'*dN;
  Fe=Ae2*fe*ones(3,1)/6;
end
```

2.4.3.2 Assembly

It only remains to **assemble** the element contributions and impose the Dirichlet boundary conditions to end up with the global stiffness matrix and load vector.

Here we will use a trick that simplifies the coding: Instead of taking care of the boundary conditions during the assembly procedure, we will *assemble the global matrix and load vector as if there were no boundary conditions, and then correct the lines that correspond to unknowns with imposed value*. This is a convenient alternative to the algorithm in §1.4.4 because Octave/MATLAB can efficiently insert/add a submatrix into a larger matrix.

If this trick is adopted, the assembly procedure of finite element stiffness matrices and load vectors is quite intuitive and straightforward to code. In Octave/MATLAB it reads:

```
LG=LV; nod=size(X,2); nel=size(LG,2); npe=size(LG,1);
K=zeros(nod,nod); F=zeros(nod,1);
for iel=1:nel
%% setting the local data
  lge=LG(:,iel);
  xe(1:dd,1:npe)=X(1:dd,lge(1:npe));
  ke=difcoeff(iel);
  fe=source(iel);
%% computing element K and F
  [Ke Fe]=elementKandF(xe,ke,fe);
%% assembly, from local to global
  K(lge,lge)=K(lge,lge)+Ke;
  F(lge)=F(lge)+Fe;
end
```

Here, arrays `difcoeff` and `source` contain the element-wise values k_e and f_e of k and f, respectively. Array `lge` is an index array that contains the numbering of the three nodes of the element. In this way, the matrix `K(lge,lge)` is the submatrix (or "slice") of `K` consisting of just the rows and columns present in `lge`. The key assembly operations

```
  K(lge,lge)=K(lge,lge)+Ke;
  F(lge)=F(lge)+Fe;
```

are thus just a shorter way of coding

```
for j=1:3
   for k=1:3
      K(lge(j),lge(k))=K(lge(j),lge(k))+Ke(j,k);
   end
   F(lge(j))=F(lge(j))+Fe(j);
end
```

which is the lengthier code we introduced originally. For example, if `lge=[7 4 9]`, then the operation

```
K(lge,lge)=K(lge,lge)+Ke;
```

is equivalent to

```
K(7,7)=K(7,7)+Ke(1,1); K(7,4)=K(7,4)+Ke(1,2); K(7,9)=K(7,9)+Ke(1,3);
K(4,7)=K(4,7)+Ke(2,1); K(4,4)=K(4,4)+Ke(2,2); K(4,9)=K(4,9)+Ke(2,3);
K(9,7)=K(9,7)+Ke(3,1); K(9,4)=K(9,4)+Ke(3,2); K(9,9)=K(9,9)+Ke(3,3);
```

It only remains to fix the lines corresponding to the boundary nodes and we will be in a position to solve our first two-dimensional problem with finite elements! Remember what we need to do: If node A is a Dirichlet node with

imposed value g, we must modify the linear system (i.e., the arrays K and F) so that the A-th equation reads, simply, $\mathsf{U}_A = g$. This is easily coded as follows:

```
ng=length(EtaG); II=eye(nod);
for ig=1:ng
 K(EtaG(ig),:)=II(EtaG(ig),:);
 F(EtaG(ig))=GG(ig);
end
```

The code is now complete, we can solve for the unknown vector U which contains the nodal values of u_h and plot the solution.

```
U=K\F;
trisurf(LG',X(1,:),X(2,:),U)
```

The command `trisurf` plots functions defined on arbitrary conforming triangulations.

Example 2.10 A uniformly heated rod of arbitrary polygonal shape.

Consider the geometry shown in Fig. 2.12, only that now we will work with the much finer discretization shown in Fig. 2.16. Let us assume that the rod is homogeneous, with diffusion coefficient equal to 1 and heat source f equal to 10, and that the surface temperature is $g = 20$. We aim to compute the temperature distribution inside the rod as given by the finite element method with a continuous P_1 finite element space.

The whole code needed for this purpose is provided as `octavefemp1a.m` in the accompanying material and reads as shown in Table 2.1. After running it, we have computed our first 2D finite element solution! The function u_h can be seen in Figure 2.17. It provides a good estimation of the exact temperature distribution, since the mesh is quite refined. It can be used to estimate, for example, the maximum temperature in the rod, which yields

$$\max_{x\in\Omega} u_h(x) = 43.077\,.$$

Example 2.11 (The uniformly heated square rod revisited with P_1 elements) By simply changing the geometry of the previous example we can revisit the square-rod problem discussed in Example 2.7.

The results obtained with different meshes are plotted in Figure 2.18. The maximum of u_h is 20.719, 20.733 and 20.737, respectively, for the meshes in part (a), (b) and (c) of the figure. The exact maximum is 20.737. The corresponding number of elements is 68, 242 and 1054, and the number of nodes is 45, 142 and 568. The maximum absolute value of the error $u - u_h$ for each mesh is 0.0193, 0.00557 and 0.00134. We observe that the numerical solution remains stable as the mesh is refined, and converges at all points of the domain to the exact solution.

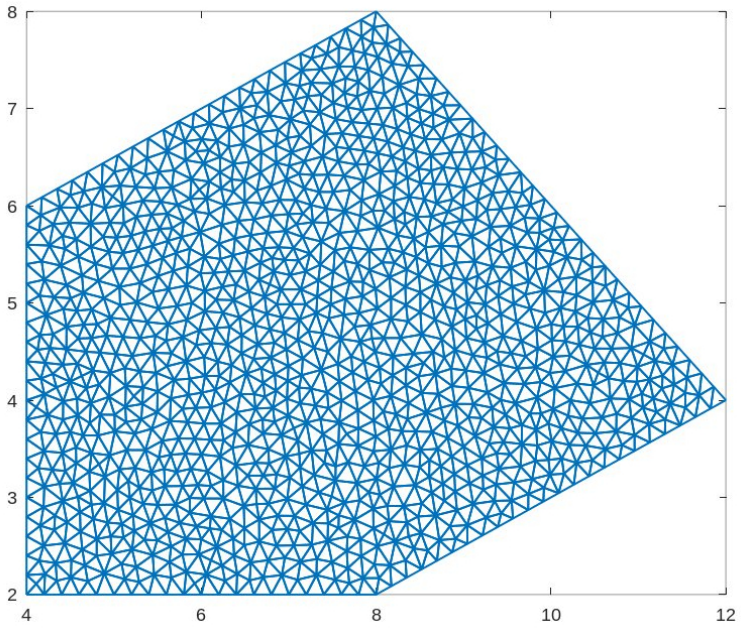


Figure 2.16 Refined triangulation of a polygonal domain.

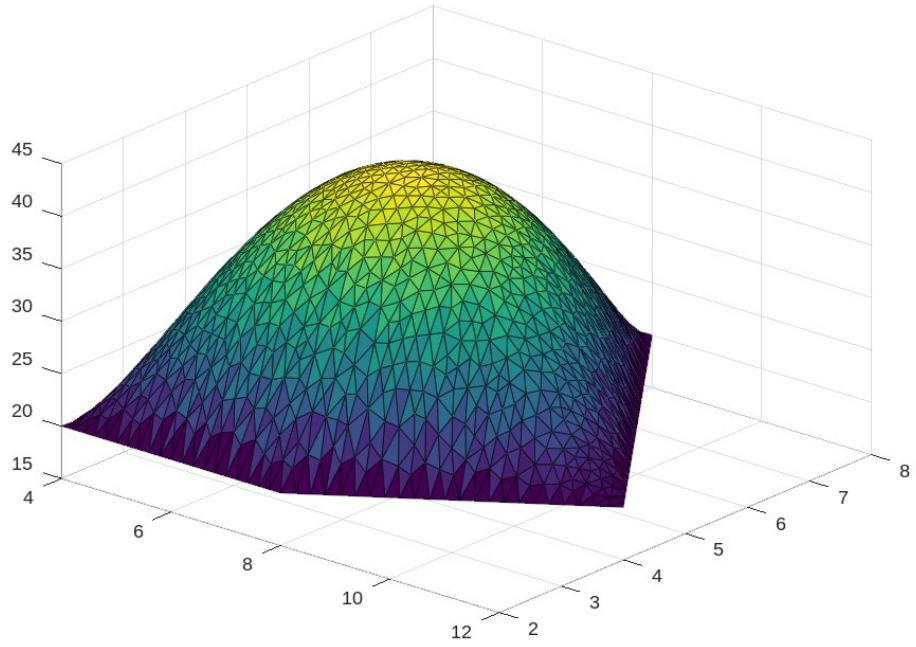


Figure 2.17 Finite element solution u_h of Example 2.10.

```
function [Ke Fe]=elementKandF(xe,ke,fe)
  dN=[xe(2,2)-xe(2,3),xe(2,3)-xe(2,1),xe(2,1)-xe(2,2);...
      xe(1,3)-xe(1,2),xe(1,1)-xe(1,3),xe(1,2)-xe(1,1)];
  Ae2=dN(2,3)*dN(1,2)-dN(1,3)*dN(2,2);
  dN=dN/Ae2;
  Ke=Ae2/2*ke*dN'*dN;
  Fe=Ae2*fe*ones(3,1)/6;
end
%% finite element solver begins (X, LV, EtaG and GG are given)
LG=LV; nod=size(X,2); nel=size(LG,2); npe=size(LG,1);
difcoeff=ones(1,nel);
source=10*ones(1,nel);
GG=20*ones(1,length(EtaG));
K=zeros(nod,nod); F=zeros(nod,1);
for iel=1:nel
%% setting the local data
  lge=LG(:,iel);
  xe(1:dd,1:npe)=X(1:dd,lge(1:npe));
  ke=difcoeff(iel);
  fe=source(iel);
%% computing element K and F
  [Ke Fe]=elementKandF(xe,ke,fe);
%% assembly, from local to global
  K(lge,lge)=K(lge,lge)+Ke;
  F(lge)=F(lge)+Fe;
end
%% boundary nodes
ng=length(EtaG); II=eye(nod);
for ig=1:ng
 K(EtaG(ig),:)=II(EtaG(ig),:);
 F(EtaG(ig))=GG(ig);
end
%% solve algebraic system
U=K\F;
%% plot
trisurf(LG',X(1,:),X(2,:),U)
```

Table 2.1 Code **octavefemp1a.m**. It solves Example 2.10.

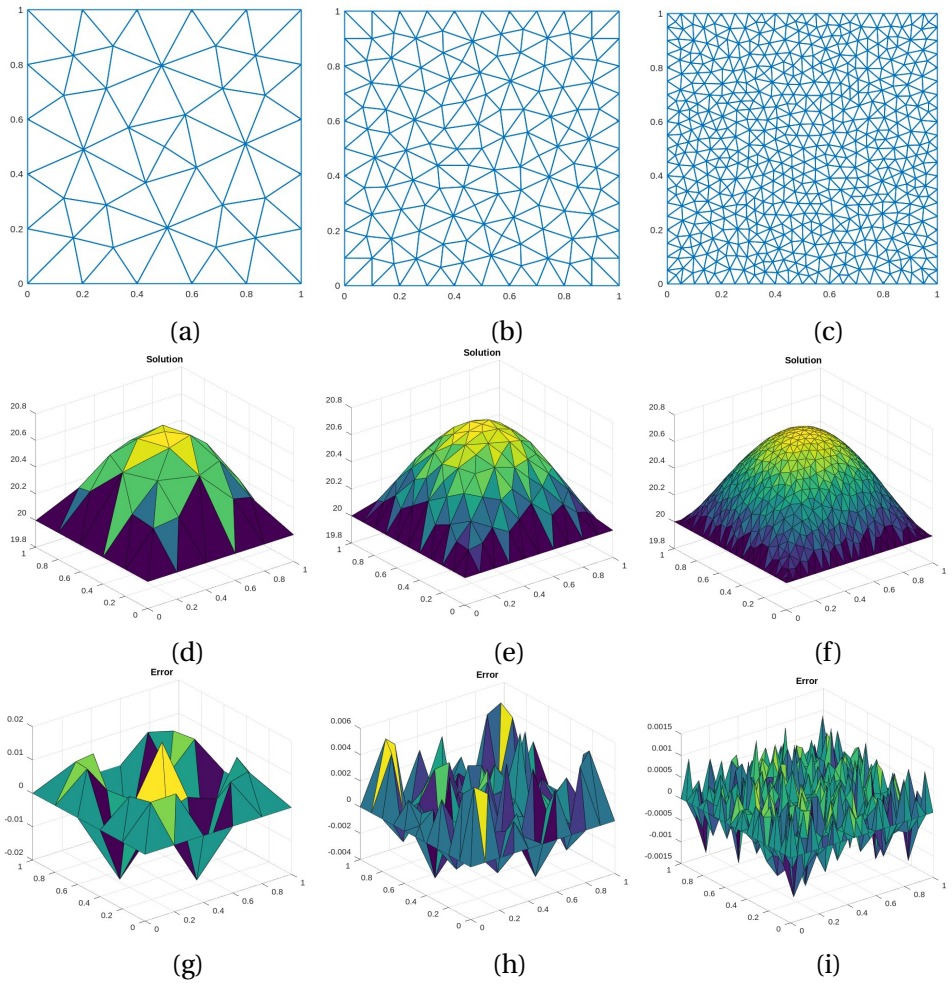


Figure 2.18 The P_1 finite element solutions of the uniformly heated square-rod example. Each column corresponds to a different mesh, which is plotted at the top of the column. The second row shows the finite element solution u_h and the third row shows the error $u - u_h$.

> For comparison purposes we report also the results from a variational method obtained, for this specific case, with the space $\mathcal{W}_h$ of quartic polynomials used in Example 2.7. The maximum of the discrete solution is 20.781 and the maximum error is 0.0445.

2.4.4 Solving Problems with Neumann Boundaries

We have presented a method that approximates the solution of the diffusion equation in any 2D polygonal domain with P_1 finite elements, but just when the solution value is known at the whole boundary (i.e., when the Neumann boundary $\partial\Omega_N$ is empty).

Frequently, however, we have information of the value of the normal flux $H = k\nabla u \cdot \check{n}$ on some parts of the boundary (where $\check{n}$ is the outward normal to $\partial\Omega$). In such cases, the solution is specified on the Dirichlet boundary as before, and the trial and test finite element spaces are still given by (2.53), but there will remain a part $\partial\Omega_N$ where there exist basis functions of $\mathcal{V}_h$ that are not identically zero.

Assuming that node A belongs to $\partial\Omega_N$, the A-th component of the load vector involves two terms,

$$\mathsf{F}_A = \int_\Omega f\, N_A\, d\Omega + \int_{\partial\Omega_N} H\, N_A\, d\Gamma\,,$$

while the corresponding line of the stiffness matrix remains the same.

Symmetry lines and homogeneous Neumann conditions. If the domain and data of the problem are symmetric with respect to some line that crosses the domain, then we can expect the solution to also be. Then one can solve the corresponding equation on a *reduced domain*, consisting of just one half of the original one, and extend the solution found in the reduced domain to the original one by symmetry. The symmetry line is thus a part of the boundary of the reduced domain. What boundary condition should be imposed there? It can be shown that the correct boundary condition is a *Neumann condition* with $H = 0$, because the normal derivative of u must be zero there. These are called **homogeneous Neumann conditions**, and also apply at any isolated (or zero-flux) boundary.

If all parts of the boundary that do not have Dirichlet conditions have homogeneous Neumann conditions, then the code in Table 2.1 works perfectly well as is. It was built considering that the Dirichlet boundary occupies the whole of $\partial\Omega$ and thus the load vector F lacks the contribution

$$\int_{\partial\Omega_N} H\, N_i\, d\Gamma,$$

but, if $H = 0$ all over $\partial\Omega_N$, **there is nothing to be added**. The variational method automatically imposes homogeneous Neumann conditions all over the part of the boundary where the solution value is not imposed. This is because homogeneous Neumann conditions are sometimes called "do-nothing boundary conditions" for this problem.

Example 2.12 (Exploiting symmetries) The case of the uniformly heated square rod exhibits several symmetries. The central horizontal and vertical lines are lines of symmetry, as well as both diagonals. By exploiting those symmetries, we end up with the reduced domain defined as the triangle ABC shown in Fig. 2.19(a). Along the edge BC the Dirichlet condition $u = g$ is applied. The other two edges (AB and CA) must satisfy **homogeneous Neumann boundary conditions** for the solutions in the original and reduced domains to coincide. This is quite useful, since meshing the reduced domain requires significantly less elements of any given size than those required by the original domain (approximately 1/8).

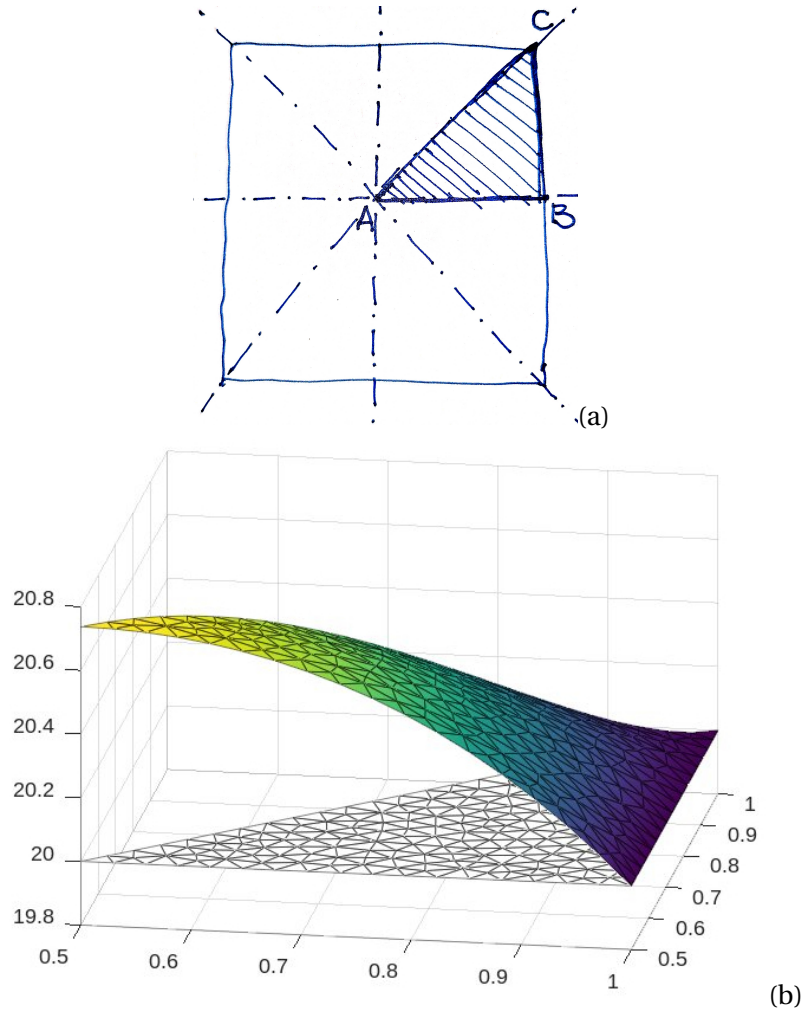


Figure 2.19 (a) Original domain of the square rod problem, and reduced domain (the triangle ABC after exploiting the several symmetries. (b) Finite element solution in the reduced domain, with homogeneous Neumann conditions at symmetry boundaries.

> We modified the previous code to set the triangle ABC as domain and only set the nodes in the edge BC as Dirichlet nodes, with imposed value $g = 20$. Nothing special was programmed for the other boundary nodes, they were treated just like internal nodes. The other constants are as in Example 2.11, i.e., $k = 1$ and $f = 10$. The result obtained with a mesh of 380 elements and 220 nodes is shown in Fig. 2.19(b). The maximum error of the discrete solution is 0.00064, smaller than the error attained in the whole domain with a mesh of 1054 elements and 568 nodes.

Non-homogeneous Neumann boundary conditions. When $H \neq 0$ one needs to build the complete load vector, which, as said, if A is a non-Dirichlet node

reads

$$\mathsf{F}_A = \int_\Omega f\, N_A\, d\Omega + \int_{\partial\Omega_N} H\, N_A\, d\Gamma\,.$$

The second integral is our focus of attention now. The differential $d\Gamma$ is a differential of length, because $\partial\Omega$ is one-dimensional. Our first task is to describe the **data structure** with which we handle the definition of $\partial\Omega_N$ and of the function H inside the code.

The **mesh generator** builds the domain boundary by joining together ℓ individual lines provided by the user, of which some (or all) belong to the Neumann boundary. Let us assume that the produced triangulation has n_{be} boundary edges. By construction, each boundary edge belongs to one and only one of these lines. The generator provides an **array of boundary edges** BE of dimension $3 \times n_{\text{be}}$. Each column of BE contains three numbers. The first two indicate the two nodes of the corresponding edge. The third one indicates the line to which the edge belongs.

To keep things simple, we can assume that H is constant over each edge. Then it can be provided by an array H of n_{be} entries, as shown in the example below.

Example 2.13 (Boundary arrays of a triangulation) In Figure 2.20 we show a small triangulation of a polygon, defined by points 1-5 and lines 1-5. For this triangulation $n_{\text{be}} = 11$, and BE is as follows:

$$\mathsf{BE} = \begin{pmatrix} 1 & 6 & 2 & 7 & 5 & 8 & 9 & 4 & 10 & 3 & 11 \\ 6 & 2 & 7 & 5 & 8 & 9 & 4 & 10 & 3 & 11 & 1 \\ 1 & 1 & 2 & 2 & 3 & 3 & 3 & 4 & 4 & 5 & 5 \end{pmatrix}$$

Now suppose that along **line 2** the imposed heat flux H is constant and equal to $\frac{2}{3}$, while along **line 3** it is also constant but equal to $\frac{1}{2}$. Along lines 1, 4 and 5 the Dirichlet condition $u = 20$ is imposed. This is specified in the array H that provides the value of H over each edge, assuming it is constant at each edge.

$$\mathsf{H} = \begin{pmatrix} 0 & 0 & \frac{2}{3} & \frac{2}{3} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \end{pmatrix}$$

The construction of H is easily coded as

```
for k=1:nbe
   if (BE(3,k)==2)
      HH(k)=2./3;
   end
   if (BE(3,k)==3)
      HH(k)=1./2;
   end
end
```

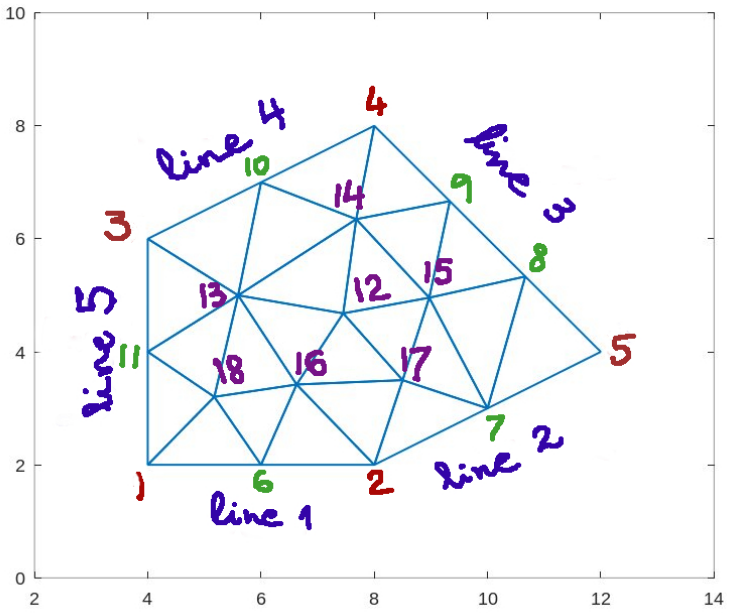


Figure 2.20 A small triangulation, showing the numbering of the defining points (in red), of the defining lines (in blue), and of the rest of the vertices.

It should be clear by now that a P_1 function restricted to an edge of the triangulation is *affine*. The function interpolates linearly along the edge between the two nodal values at its ends. This implies that, if the k-th boundary edge (let us denote it by E_k) joins the nodes $A_1 = \mathsf{BE}(1,k)$ and $A_2 = \mathsf{BE}(2,k)$, then the only two basis functions that are different from zero along the edge are N_{A_1} and N_{A_2}. Further, these two basis functions go affinely from 0 to 1 over the edge, so that their average value is $\frac{1}{2}$. We can thus compute the integral over E_k as

$$\int_{E_k} H\, N_{A_1}\, d\Gamma = \int_{E_k} H\, N_{A_2}\, d\Gamma = \frac{1}{2}\,\mathsf{H}(k)\,|E_k|\,,$$

where $|E_k|$ is the length of the edge,

$$|E_k| = \|X^{A_1} - X^{A_2}\|\,.$$

We can implement the addition of the Neumann part of the load vector as an assembly procedure.

```
%% Neumann contributions
for ied=1:nbe
  lged=BE(1:2,ied);
  xed(1:dd,1:2)=X(1:dd,lged);
  Led=norm(xed(:,1)-xed(:,2));
  Hed=HH(ied);
  F(lged)=F(lged)+Hed*Led/2;
end
```

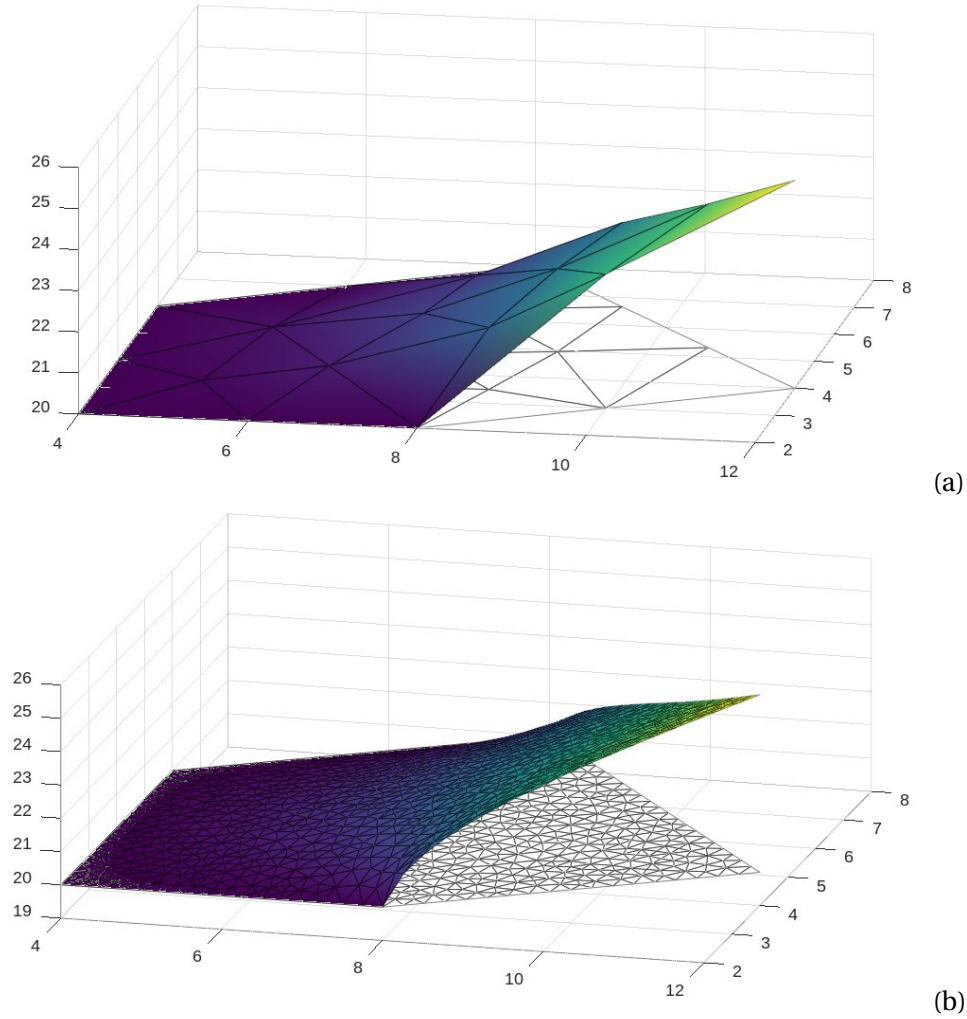


Figure 2.21 Solution to Example 2.14 with two meshes with different resolutions.

Remark: Of course, the contribution of each edge will depend on the specific function H. Piecewise polynomial functions of higher degree can easily be implemented.

Example 2.14 (Solution of the Neumann problem of Example 2.13) The procedure above has been implemented in **octavefemp1b.m**. The source f was set to zero and the diffusion coefficient to $q = 1$. The results obtained for the mesh of Figure 2.20 and for a finer mesh on the same geometry are shown in Figure 2.21.

Chapter 3

Numerical Analysis of the FEM for Elliptic Problems

3.1 A Short Recapitulation

Let u be the **exact solution** of the problem of our interest, posed over a domain Ω. For our analysis we can assume u to be smooth (e.g., C^∞) and thus a solution of the corresponding strong form. The problem can be a second or fourth order elliptic problem in one or more dimensions, with suitable boundary conditions.

In the previous chapters, by multiplying the differential equation by an arbitrary smooth function v and integrating over Ω, we arrived at a bilinear form $a(\cdot,\cdot)$ and a linear form $\ell(\cdot)$ such that

$$a(u,v) = \ell(v)\ , \tag{3.1}$$

which holds for any smooth v that satisfies **homogeneous essential boundary conditions**. In other words, v satisfies the **essential boundary conditions** with value **zero**. The identification of which boundary conditions are essential was part of the process of deriving (3.1).

Our next step was to introduce a **finite element space** $\mathcal{W}_h$ defined on a mesh over Ω. The space $\mathcal{W}_h$ was in fact introduced by providing a set of **basis functions**, $\{\mathcal{N}_a, a = 1,2,\ldots,n\}$, such that

$$w_h \in \mathcal{W}_h \quad \Leftrightarrow \quad w_h(x) = \sum_{a=1}^{n} c_a N_a(x)\ .$$

This is the way finite element spaces are specified in practice.

We defined the **trial space** $\mathcal{S}_h$ and the **test space** $\mathcal{V}_h$ as

$$\mathcal{S}_h = \{w_h \in \mathcal{W}_h \mid w_h \text{ satisfies essential boundary conditions}\}$$
$$\mathcal{V}_h = \text{Direction of } \mathcal{V}_h.$$

Remark: The possibility of taking $\mathcal{S}_h$ and $\mathcal{V}_h$ totally independent of one another certainly exists and has been explored in the literature (we called the Petrov-Galerkin methods). The methods we consider in this book, for which $\mathcal{V}_h$ is the

direction of $\mathcal{S}_h$ and both are contained in an encompassing $\mathcal{W}_h$, remain the most popular choice.

Remark: When the domain is one-dimensional the essential boundary conditions consist of the value of u and/or u' at the boundary points. They can always be imposed exactly. In two or more dimensions the boundary values are arbitrary functions specified over lines or surfaces. If the boundary values are not piecewise polynomials, we assume that $\mathcal{S}_h$ satisfies an approximate (interpolated) version of the boundary values. It can be rigorously justified that the approximation of the boundary values does not hinder the convergence of the method. We will come back to this issue later on.

The last step in our construction was to define the **finite element solution** $u_h \in \mathcal{S}_h$ by

$$a(u_h, v_h) = \ell(v_h) \,, \qquad \forall\, v_h \in \mathcal{V}_h \,. \tag{3.2}$$

At this point, other than knowing that u_h and u satisfy (at least approximately) the same essential boundary conditions, we only know that u_h satisfies, over $\mathcal{V}_h$, the same variational equation that u satisfies for **smooth** functions (functions in $\mathcal{V}$). Is this sufficient to guarantee that u_h approximates u in some sense? What are the conditions on $\mathcal{W}_h$, $a(\cdot,\cdot)$ and $\ell(\cdot)$ for u_h to converge to u as the mesh is refined?

These questions concern the **numerical analysis** of the FEM, which deals with the mathematical properties of the method and is the subject of this chapter. We will make a brief excursion into the most important theorems and definitions, so as to provide a basis for understanding the specialized literature.

3.2 The Fundamental Approximation Result

The following is the fundamental theorem on which most of the analysis is based.

Theorem 3.1. *(***Céa's Lemma***) Assume that the exact solution satisfies the discrete problem (***exact consistency***):*

$$a(u, v_h) = \ell(v_h)\,, \qquad \forall\, v_h \in \mathcal{V}_h\,. \tag{3.3}$$

Assume, further, that there is a **norm** $\|\cdot\|$ *(defined on functions $f : \Omega \to \mathbb{R}$), such that:*

1. **Domain of the Norm:** $\|u\| < +\infty$, $\|w_h\| < +\infty$, $\forall\, w_h \in \mathcal{W}_h$.
2. **Continuity:** *There exist $M > 0$ and $m > 0$ such that*

$$|a(u - w_h, v_h)| \le M\,\|u - w_h\|\,\|v_h\|, \qquad \forall\, v_h \in \mathcal{V}_h, \forall\, w_h \in \mathcal{S}_h \tag{3.4}$$

$$|\ell(v_h)| \le m\,\|v_h\|, \qquad \forall\, v_h \in \mathcal{V}_h \tag{3.5}$$

3. **Coercivity:** *There exists $\alpha > 0$ such that*

$$a(v_h, v_h) \ge \alpha\,\|v_h\|^2\,, \qquad \forall\, v_h \in \mathcal{V}_h\,. \tag{3.6}$$

Then,

a) *the finite element solution* **exists**, *is* **unique**, *and satisfies the* **stability** *estimate*

$$\|u_h\| \le \frac{m}{\alpha} + \left(1 + \frac{M}{\alpha}\right) \min_{w_h \in \mathcal{S}_h} \|w_h\|\,; \tag{3.7}$$

b) *the following* **a priori approximation result** *holds:*

$$\|u - u_h\| \le \left(1 + \frac{M}{\alpha}\right) \min_{w_h \in \mathcal{S}_h} \|u - w_h\| \tag{3.8}$$

Let us briefly recall what a **norm** is.

Definition 3.1 (Norm)**.** *Let V be a vector space. A norm is a function $\|\cdot\| : V \to \mathbb{R}$ such that for $v, u \in V$ and $\beta \in \mathbb{R}$:*

1. **N.1.** $\|v\| \ge 0$, *and* $\|v\| = 0$ *if and only if* $v = 0$.
2. **N.2.** $\|\beta v\| = |\beta|\,\|v\|$.
3. **N.3.** $\|v + u\| \le \|v\| + \|u\|$ *(triangle inequality).*

The typical norm that you are familiar with is the "Euclidean norm" in $\mathbb{R}^3$: If $x = (x_1, x_2, x_3)$, then $\|x\| = \sqrt{x_1^2 + x_2^2 + x_3^2}$. Clearly, N.1 holds, since $\|x\| \ge 0$, and if $\|x\| = 0$, then $x = 0$. The second condition, N.2, is also simple to verify, and the triangle inequality is the common statement that the sum of the lengths of two sides of a triangle is always greater or equal than the length of the third. These three conditions are intuitive to understand in the case of $\mathbb{R}^n$, and the fact that the Euclidean norm satisfies them is easy to see. Defining a norm for vector spaces of

functions and proving N.1-N.3 is a little bit more complex, but can be done. Some such norms will be introduced soon. A more extensive discussion with examples can be found in §A.

Going back to Thm. 3.1, it is easy to see why this theorem, assuming its hypotheses are met, provides useful information about the numerical solution u_h:

- **Conclusion a)** guarantees that u_h exists and is unique. Since u_h is defined algorithmically as the solution of the algebraic system

$$\text{stiffness matix} \quad \times \quad \text{unknown vector} \quad = \quad \text{load vector}$$

  this conclusion guarantees that the stiffness matrix is non-singular.

- Further, this conclusion tells us what to expect of the norm of u_h as the mesh is refined. Assume that α does not depend on the mesh size h, and also accept the technical hypothesis that the minimum in (3.7) is bounded independently of h. Then there exists a constant C such that

$$\|u_h\| \leq C$$

  for all mesh sizes $h > 0$. The discrete solution is uniformly bounded (in the norm $\|\cdot\|$ of course) for arbitrarily refined meshes.

  Remark: The value of $A_h = \min_{w_h \in \mathcal{S}_h} \|w_h\|$ depends on how "nice" the essential boundary conditions are. If, for example, the boundary is a curve (2D problem) and the imposed boundary value is discontinuous along the curve, then A_h could (or will) grow to infinity as the mesh is refined. If the boundary data are continuous then A_h remains bounded as $h \to 0$.

  Remark: In a first reading, it may prove useful to consider that the essential boundary conditions are zero. This makes $\mathcal{S}_h = \mathcal{V}_h$ and thus $0 \in \mathcal{S}_h$, implying that $A_h = 0$.

- **Conclusion b)** tells us how the error $u - u_h$ behaves. Let us define a function $w_h^* \in \mathcal{S}_h$ that realizes the minimum in (3.8). This function is thus, as measured by the norm $\|\cdot\|$, the **closest approximation** to u from the finite element space $\mathcal{S}_h$. It is computable *if the exact solution u is known* by solving a minimization problem.

  We cannot expect (though in some cases it happens!) that the finite element solution, which is computed *without knowing u*, coincides with the best possible approximation w_h^*. But conclusion b) tells us that $\|u - u_h\|$ is **at most** a factor $(1 + M/\alpha)$ larger that $\|u - w_h^*\|$.

  Further, **if the constants** M **and** α **do not depend on** h, we have that

$$\text{if } \|u - w_h^*\| \to 0 \text{ as the mesh is refined} \quad \Rightarrow \quad \|u - u_h\| \to 0$$

In other words: "If the finite element trial space **can** approximate the exact solution, it (eventually) **will**".

Mathematically speaking, one can create a **sequence** of successively refined meshes of sizes $h_k \to 0$, obtain the corresponding FEM solutions u_{h_k}, and this sequence will **converge** to the exact solution u in the norm $\|\cdot\|$.

As we see, if a finite element method satisfies the hypotheses of Theorem 3.1 we have substantial guarantees with respect to its numerical performance. The mathematical analysis of a method usually begins by determining the appropriate norm and checking the hypotheses of the Theorem. Though the details depend on the specific problem and norm, it is worth making some general comments about each of the hypotheses.

Domain of the Norm. To be able to use the norm to measure distances among the exact solution u and/or any of the functions in the approximation space $\mathcal{W}_h$, the norm has to be defined on u and each function $w_h \in \mathcal{W}_h$. This is what this hypothesis requires.

Exact consistency. The consistency of the discrete formulation depends on the adopted variational equation and on the finite element space $\mathcal{W}_h$. In the previous chapters we have seen finite element formulations for second and fourth-order elliptic problems in one or more dimensions. For each formulation, we introduced a finite element space that "works," and we verified that it works properly through numerical tests. For **second order problems** we selected finite element spaces consisting of C^0 functions, while for **fourth order problems** they consisted of C^1 functions. As we have seen, the main reason for this is that by enforcing such continuity on $\mathcal{W}_h$ we obtain a method that satisfies (3.3).

Condition (3.3) can be equivalently expressed as the so-called **Galerkin orthogonality**,

$$a(u - u_h, v_h) = 0\,, \qquad \forall\, v_h \in \mathcal{V}_h\,. \tag{3.9}$$

To check for consistency, one considers the **consistency residual**

$$r(u, v_h) = a(u, v_h) - \ell(v_h)\,.$$

If it is identically zero, as said, the method is **exactly consistent**. If it is not identically zero, then the theorem cannot be applied and, in general, u_h does not converge to u. There are exceptions, however. **Approximately consistent** formulations can still yield convergent results, but they are outside of our scope at this point.

Continuity. Inequality (3.4) is the definition of the continuity of a bilinear form, with respect to both its arguments. The continuity with respect to the first argument, in particular, guarantees that, if $\|u - w_h\|$ is small, then $a(u - w_h, \cdot)$ is also

small. It can also be rewritten as

$$|a(u, v_h) - a(w_h, v_h)| \leq M\|u - w_h\| \, \|v_h\| \, ,$$

which brings forth that, if $\|u - w_h\|$ is small, then the difference between $a(u, \cdot)$ and $a(w_h, \cdot)$ is also small, as applied on any v_h.

Similarly, (3.5) implies that $|\ell(v_h) - \ell(w_h)| = |\ell(v_h - w_h)| \leq m\|v_h - w_h\|$ for all w_h and v_h in $\mathcal{W}_h$, which again shows that if $\|v_h - w_h\|$ is small then the difference between $\ell(v_h)$ and $\ell(w_h)$ is small.

In fact, since $\mathcal{W}_h$ is finite dimensional, such constants M and m always exist (because linear or bilinear functions are always continuous in finite dimensional spaces). However, in most cases of interest the continuity hypothesis holds in a **uniform** sense, i.e., with M and m **independent of the mesh size** h.

Coercivity. Inequality (3.6) defines the so-called **strong coercivity** of the bilinear form $a(\cdot, \cdot)$ over the space $\mathcal{V}_h$. It implies that

$$a(v_h, v_h) > 0 \, , \qquad \forall \, v_h \in \mathcal{V}_h, \; v_h \neq 0 \, . \tag{3.10}$$

Testing the strong coercivity of a bilinear form is thus easy, one simply plugs a generic v_h of the test space in both arguments of $a(\cdot, \cdot)$ and checks that the result is zero if and only if $v_h = 0$. If there exists $v_h \neq 0$ such that $a(v_h, v_h) = 0$, then (3.6) does not hold and the theorem cannot be applied.

However, (3.10) being true does not automatically imply that (3.6) holds **with a constant α independent of h**. This condition needs to be checked mathematically and in fact guides the selection of the appropriate norm $\|\cdot\|$.

It is now time to apply Theorem 3.1 to the different finite element methods that we built in the previous chapter. This is done in the following sections. Below we provide a proof of Céa's lema for the sake of completeness.

Proof. (Proof of Thm. 3.1) Let w_h be an arbitrary function belonging to $\mathcal{S}_h$ and let $u_h \in \mathcal{S}_h$ satisfy

$$a(u_h, v_h) = \ell(v_h) \, , \qquad \forall \, v_h \in \mathcal{V}_h \, . \tag{3.11}$$

Noticing that $u_h - w_h$ belongs to $\mathcal{V}_h$, we have from the coercivity hypothesis that

$$\|u_h - w_h\|^2 \quad \leq \quad \frac{1}{\alpha} \, a(u_h - w_h, u_h - w_h)$$

and, from the definition of u_h,

$$a(u_h - w_h, u_h - w_h) = \ell(u_h - w_h) - a(w_h, u_h - w_h) \, .$$

Combining the two equations above and using the hypotheses of continuity of $a(\cdot, \cdot)$ and $\ell(\cdot)$ we get

$$\|u_h - w_h\|^2 \leq \frac{1}{\alpha} \, (m\|u_h - w_h\| + M\|w_h\| \, \|u_h - w_h\|)$$

and thus

$$\|u_h - w_h\| \le \frac{m}{\alpha} + \frac{M}{\alpha}\|w_h\|, \quad \forall\, w_h \in \mathcal{S}_h\,.$$

Applying the triangle inequality $\|u_h\| \le \|u_h - w_h\| + \|w_h\|$ and taking the minimum over $\mathcal{S}_h$ completes the proof of (3.7).

In fact, (3.7) implies the uniqueness of u_h. To see this, notice first that u_h results from a *square linear system of equations*

$$\mathsf{K}\mathsf{U} = \mathsf{F} \tag{3.12}$$

which has the dimension of $\mathcal{W}_h$. From linear algebra we know that, if the solution is not unique, there exists a vector Z such that $\mathsf{U} + \beta\mathsf{Z}$ satisfies the linear system, for all $\beta \in \mathbb{R}$. Alternatively, there exists a non-zero vector Z such that $\mathsf{K}\mathsf{Z} = 0$.

We have already discussed at length the construction of the linear system (3.12) starting from the weak form and a basis $\{N_a\}$ of $\mathcal{W}_h$. It is clear that, if $\mathsf{U} + \beta\mathsf{Z}$ satisfies (3.12), then $u_h + \beta z_h$ belongs to $\mathcal{S}_h$ and satisfies (3.11), where $z_h = \sum_a Z_a N_a$.

We can thus apply inequality (3.7), that we already proved, to infer that

$$\|u_h + \beta z_h\| \le C\,, \qquad \forall\, \alpha \in \mathbb{R}$$

with $C = m/\alpha + (1 + M/\alpha)\min_{w_h}\|w_h\|$. But this can only be satisfied for all β if $z_h = 0$,[1] which implies $\mathsf{Z} = 0$ and thus that U and u_h are unique.

To finish the proof of a) it only remains to prove the existence of U (and thus of u_h). However, it is known from elementary linear algebra (the rank-nullity theorem) that for square systems *uniqueness implies existence*. Alternatively, if the system $\mathsf{K}\mathsf{Z} = 0$ has only the trivial solution $\mathsf{Z} = 0$, then K is invertible and the solution exists and is computed as $\mathsf{U} = \mathsf{K}^{-1}\mathsf{F}$.

Let us now turn to prove conclusion b). Again, let w_h be an arbitrary element of $\mathcal{S}_h$. Then,

$$\begin{aligned}
\alpha\|u_h - w_h\|^2 &\le a(u_h - w_h, u_h - w_h) && \text{coercivity, (3.6)}\\
&= \underbrace{a(u_h - u, u_h - w_h)}_{=0} + a(u - w_h, u_h - w_h) && \text{add and subtract } u\\
&= a(u - w_h, u_h - w_h) && \text{Galerkin orthogonality, (3.9)}\\
&\le M\|u - w_h\|\|u_h - w_h\| && \text{continuity, (3.4).}
\end{aligned}$$

Notice that it is possible to use Galerkin orthogonality above because $w_h - u_h \in \mathcal{V}_h$, or because $\mathcal{V}_h$ is the direction of $\mathcal{S}_h$. We can then conclude that

$$\|u_h - w_h\| \le \frac{M}{\alpha}\|u - w_h\|.$$

[1] Specifically, by the triangle inequality, $\beta\|z_h\| \le \|u_h + \beta z_h\| + \|-u_h\| \le C + \|-u_h\|$, and hence $0 \le \|z_h\| \le (C + \|-u_h\|)/\beta$ for any β. This means that $\|z_h\|$ is smaller than any positive number by selecting β large enough, and hence $z_h = 0$.

Application of the triangle inequality leads to (3.8) because, for all w_h,

$$\begin{aligned}\|u-u_h\| &\le \|u-w_h\| + \|w_h-u_h\| \\ &\le \|u-w_h\| + \frac{M}{\alpha}\|u-w_h\| = \left(1+\frac{M}{\alpha}\right)\|u-w_h\|.\end{aligned}$$

□

3.3 Second Order Problems in One Dimension

Let us proceed to show how Theorem 3.1 can be applied to analyze finite element methods for second order elliptic problems in one dimension.

Let $u:[0,L]\to\mathbb{R}$ be the exact solution of Problem 1.1 with $b=0$, i.e., u satisfies $u(0)=g_0$, $u'(L)=d_L$ and

$$-\big(k(x)u'(x)\big)' + c(x)u(x) = f(x)\,, \qquad \forall\, x\in(0,L)\,. \tag{3.13}$$

Physical considerations allow us to consider that the coefficients k and c satisfy the bounds

- From below: $k(x)\ge k_{\min}>0$ and $c(x)\ge 0$.
- From above: $k(x)\le k_{\max}<+\infty$ and $c(x)\le c_{\max}<+\infty$.

For $f(x)$ we assume that it is square-integrable, i.e., that the L^2-norm of f

$$\|f\|_0 = \left(\int_0^L f(x)^2\,dx\right)^{\frac{1}{2}} \tag{3.14}$$

is finite.

A variational equation for this problem was obtained in §1.1.2.3. It was shown that

- the corresponding bilinear form and linear functional are

$$a(u,v)=\int_0^L (ku'v'+cuv)\,dx\,, \qquad \ell(v)=\int_0^L fv\,dx \;+\; k(L)d_L v(L)\,, \text{ and}$$

- the **essential** boundary condition is $u(0)=g_0$, while the condition $u'(L)=d_L$ is a **natural** boundary condition (it is incorporated into $\ell(v)$, see above).

To approximate u, we choose a finite element space $\mathcal{W}_h$ from which we define the trial and test spaces as

$$\mathcal{S}_h=\{w_h\in\mathcal{W}_h \,|\, w_h(0)=g_0\}\,, \qquad \mathcal{V}_h=\{w_h\in\mathcal{W}_h \,|\, w_h(0)=0\}\,.$$

and then compute the **finite element solution** $u_h\in\mathcal{S}_h$ by solving

$$a(u_h,v_h)=\ell(v_h)\,, \qquad \forall\, v_h\in\mathcal{V}_h\,. \tag{3.15}$$

We will assume that functions in the finite element space $\mathcal{W}_h$ are C^1 in each closed element. We now study whether u_h indeed provides an approximation of u that becomes closer and closer as the mesh is refined.

3.3.1 Approximation Result

Let us analyze the validity of each one of the hypotheses of Theorem 3.1.

3.3.1.1 Exact Consistency

The consistency residual is given by

$$r(u,v_h)=\int_0^L (ku'v_h'+cuv_h)dx-\int_0^L fv_h dx-k(L)d_L v_h(L)\,.$$

The exact solution u is assumed to be a smooth function, and v_h by construction is smooth inside each element K in the mesh $\mathcal{T}_h$. We can thus integrate by parts (Theorem 1.2) in each K to get

$$\begin{aligned} r(u,v_h) &= \int_0^L [-(ku')'+bu'+cu-f]\,v_h\,dx \\ &\quad +k(L)(u'(L)-d_L)\,v_h(L) \\ &\quad +k(0)u'(0)\,v_h(0) \\ &\quad +\sum_z k(z)u'(z)\,(v_h(z^-)-v_h(z^+))\,, \end{aligned}$$

where z runs over all interelement boundaries and $v_h(z^-)-v_h(z^+)$ is the jump of v_h across z.

Remembering that u is the exact solution to (3.13), one automatically has that the first and second terms of the consistency residual are zero.

The third and fourth terms, on the other hand, are not guaranteed *by the solution* u to be identically zero because in general u' is not zero at $x=0$ or at the element boundaries. Thus, **for exact consistency to hold automatically, the space $\mathcal{V}_h$ must consist of continuous functions that are zero at $x=0$.** This makes $v_h(0)$ and the jumps in v_h identically zero, and thus $r(u,v_h)=0$ for all v_h.

The satisfaction of the consistency condition requires finite element spaces $\mathcal{V}_h$ that consist of continuous functions that are zero at the essential boundary. Because $\mathcal{V}_h$ is the direction of $\mathcal{S}_h$, it is convenient to set $\mathcal{S}_h$ to be continuous as well , or directly, to build a space $\mathcal{W}_h$ of continuous functions.

This results is a generalization of the results in §1.4.1.1, in a little more detail.

Even if $\mathcal{V}_h$ contains only continuous functions, $\mathcal{S}_h$ could have *all* discontinuous functions if, for example, we set $\mathcal{S}_h=\{w_h=v_h+s_h \mid v_h\in\mathcal{V}_h\}$ where s_h is a discontinuous function.

3.3.1.2 Coercivity and the H^1-Norm

We look for a norm such that there exists $\alpha>0$ (preferrably independent of h) satisfying

$$a(v_h,v_h)=\int_0^L [k(x)v_h'(x)^2+c(x)v_h(x)^2]\,dx\geq\alpha\|v_h\|^2 \tag{3.16}$$

for all $v_h\in\mathcal{V}_h$. Indeed, we will prove that a norm that satisfies this is the so-called H^1-norm,

$$\|v_h\|_1=\left[\int_0^L \left(v_h'(x)^2+v_h(x)^2\right)\,dx\right]^{\frac{1}{2}} = \left(\|v_h\|_0^2+\|v_h'\|_0^2\right)^{\frac{1}{2}}. \tag{3.17}$$

The proof is very easy if we add the hypothesis $c(x) \geq c_{\min} > 0$ because in such a case

$$a(v_h, v_h) = \int_0^L [k(x) v_h'(x)^2 + c(x) v_h(x)^2] \, dx$$
$$\geq \min\{k_{\min}, c_{\min}\} \underbrace{\int_0^L [v_h'(x)^2 + v_h(x)^2] \, dx}_{\|v_h\|_1^2}$$

and (3.16) holds with $\alpha = \min\{k_{\min}, c_{\min}\}$. Notice that α is totally independent of the mesh.

However, it is not unfrequent to deal with problems in which $c(x)$ is either zero everywhere (a pure diffusion problem) or at some region of the domain. In such a case coercivity in the H^1-norm still holds, but the proof is slightly different and relies on v_h being zero at $x = 0$ (which was not used above). Details are given as complementary material.

Coercivity is a very strong condition for a variational equation. It is not necessary for convergence, since there can be convergence without coercivity, but its absence may indicate issues with either the numerical method, the exact problem being solved, or the variational equation used. For example, if we tried to approximate a function u for which all we know is that

$$\mu(u(0) - g) v(0) = 0$$

with $\mu > 0$, one of the variational equations we used to formulate Nitsche's method in one dimension, we would immediately find that it is (generally) not coercive. In this case the bilinear form is $a(u, v) = \mu u(0) v(0)$, and $a(v, v) = 0$ for any v such that $v(0) = 0$ in any test space that contains a non-zero function that is equal to 0 at $x = 0$. This lack of coercivity is a reflection of the fact that the exact problem being solved is not well-posed: there are many functions u that are equal to g at $x = 0$, so the variational equation does not uniquely define u. The variational method inherits this feature, and u_h does not have a unique solution to which it should converge to.

3.3.1.3 Domain of the Norm

Once coercivity has been verified, we need to check that the norm is finite for u and for any $w_h \in \mathcal{W}_h$. Fortunately, this is immediate, since it was assumed that u is smooth (in the *closed* interval $[0, L]$) and that w_h is continuous in $[0, L]$ and C^1 in each close element. The latter guarantees that the derivatives are continuous up to the boundary of each element, and hence bounded, so the $\|\cdot\|_1$ is finite.

3.3.1.4 Continuity

It only remains to prove continuity of $a(\cdot,\cdot)$ and $\ell(\cdot)$ in the H^1-norm.

$$
\begin{aligned}
|a(u-w_h, v_h)| &= \\
&= \left| \int_0^L \left[k(x)(u'(x) - w_h'(x))\, v_h'(x) + c(x)(u(x) - w_h(x))\, v_h(x) \right] dx \right| \\
&\le \int_0^L \left| k(x)(u'(x) - w_h'(x))\, v_h'(x) + c(x)(u(x) - w_h(x))\, v_h(x) \right| dx && \text{Triangle inequality for integrals} \\
&\le \int_0^L \left(k(x)|u'(x) - w_h'(x)||v_h'(x)| + c(x)|u(x) - w_h(x)||v_h(x)| \right) dx && \text{Triangle inequality} \\
&\le k_{\max} \int_0^L |u'(x) - w_h'(x)||v_h'(x)| dx + c_{\max} \int_0^L |u(x) - w_h(x)||v_h(x)| dx \\
&\le k_{\max} \| u' - w_h' \|_0 \| v_h' \|_0 + c_{\max} \| u - w_h \|_0 \| v_h \|_0 && \text{Cauchy-Schwartz, (3.19)} \\
&\le M \| u - w_h \|_1 \| v_h \|_1 . && \|g\|_0, \|g'\|_0 \le \|g\|_1 \,, \forall g
\end{aligned}
$$

with $M = k_{\max} + c_{\max}$.

To prove the continuity of the linear form, on the other hand, we consider

$$
\ell(v_h) = \int_0^L f(x)\, v_h(x)\, dx + k(L)\, d_L\, v_h(L) \,. \tag{3.18}
$$

The first term is easily proved to be bounded in the H^1-norm when $f \in L^2([0,L])$, since

$$
\int_0^L f(x)\, v_h(x)\, dx \le \| f \|_0 \, \| v \|_0 \le \| f \|_0 \, \| v \|_1 \,.
$$

Here we have used the **Cauchy-Schwartz inequality for integrals** which states that, if two functions v and w have finite L^2-norm, then

$$
\int_0^L v(x) w(x)\, dx \le \left(\int_0^L v^2(x)\, dx \right)^{\frac{1}{2}} \left(\int_0^L w^2(x)\, dx \right)^{\frac{1}{2}} = \| v \|_0 \, \| w \|_0 \,. \tag{3.19}
$$

The integral $\int_0^L v(x) w(x)\, dx$ is referred to as the L^2 **inner product** of v and w.

The second term in the right-hand side of (3.18) is trickier but still can be proved to be continuous without using that $v_h(0) = 0$ (so that it holds for Neumann boundary conditions too). We use the fact that v_h is *continuous* in space. As a consequence, there exists a point $\overline{x} \in [0,L]$ such that $v_h(\overline{x})$ equals the mean of v_h, i.e.,

$$
v_h(\overline{x}) = \frac{\int_0^L v_h(x)\, dx}{L} \,.
$$

From this,

$$\begin{aligned} v_h(L) &= v_h(\overline{x}) + \int_{\overline{x}}^{L} v_h'(x)\,dx \\ &= \frac{1}{L}\int_0^L v_h(x)\,dx + \int_{\overline{x}}^{L} v_h'(x)\,dx \\ &\leq \frac{1}{L}\|1\|_0\,\|v_h\|_0 + \int_0^L |v_h'(x)|\,dx \\ &\leq \frac{1}{L}\|1\|_0\,\|v_h\|_0 + \|1\|_0\|v_h'\|_0 \\ &\leq \left(\frac{1}{\sqrt{L}} + \sqrt{L}\right)^{\frac{1}{2}} \|v_h\|_1\,, \end{aligned}$$

where we have used that $\|1\|_0 = \sqrt{L}$. We have thus proved that, if f is square-integrable and L is finite, then $\ell(\cdot)$ is continuous in the H^1-norm, i.e.,

$$|\ell(v_h)| \leq m\,\|v_h\|_1\,, \qquad \forall\, v_h \in \mathcal{W}_h\,,$$

where $m = \|f\|_0 + k(L)\,d_L\left(\frac{1}{\sqrt{L}} + \sqrt{L}\right)^{\frac{1}{2}}$. **It is again important to notice that both M and m do not depend on the mesh.**

We have thus checked all the hypothesis of Thm. 3.1 and can thus infer the following theorem:

Theorem 3.2. *If the finite element space $\mathcal{W}_h$ consists of* **continuous functions** *that in addition are C^1 in each closed element, then the finite element solution u_h defined by* (3.15) *exists, is unique, and satisfies*

$$\|u - u_h\|_1 \leq C \min_{w_h \in \mathcal{S}_h} \|u - w_h\|_1 \tag{3.20}$$

with the constant C independent of the adopted mesh.

3.3.2 Convergence of the Finite Element Solution

We now know that (3.20) holds for the finite element solution of (3.13). The remaining question is: How does the right-hand side of (3.20) behave as the mesh is refined?

Let us consider the so-called P_k **Lagrange elements** in one dimension. Some of these elements were introduced in Chapter 1. The basis functions are C^0-continuous across the element boundaries, and inside each element K the basis functions span the space $P_k(K)$. The "simplest" finite element (cf. 1.4.1) space corresponds to the P_1 Lagrange element.

The P_k Lagrange elements (P_k-elements for short in what follows) are the most popular finite elements for second order elliptic problems in 1D.

A well known result of numerical analysis is the following **interpolation estimate**:

Theorem 3.3. *Let u be a C^{k+1} function in $\Omega = [0, L]$ such that $u(0) = g_0$ and let $\mathcal{W}_h$ be the P_k Lagrange finite element space built on a mesh $\mathcal{T}_h$ of Ω. Let h denote the length of the largest element in $\mathcal{T}_h$. Then, there exists a constant C_I that depends on k but not on h or u such that*

$$E_1(\mathcal{S}_h, u) = \min_{w_h \in \mathcal{S}_h} \|u - w_h\|_1 \leq C_I \, h^k \, \|u^{(k+1)}\|_0 \,, \tag{3.21}$$

and

$$E_0(\mathcal{S}_h, u) = \min_{w_h \in \mathcal{S}_h} \|u - w_h\|_0 \leq C_I \, h^{k+1} \, \|u^{(k+1)}\|_0 \,, \tag{3.22}$$

where $u^{(k+1)} = d^{k+1}u/dx^{k+1}$. The minima of $\|u - w_h\|$ in the H^1 and L^2 norms, denoted by E_1 and E_0 above, will be referred to as **interpolation errors** *of the finite element space $\mathcal{S}_h$ in the respective norms.*

From the theorem we observe that:

- The order of the interpolation error is $\mathcal{O}(h^k)$ in the H^1-norm, assuming that $\|u^{(k+1)}\|_0 < \infty$.
- If the norm does not contain the derivative, the interpolation error for the function decreases more rapidly with h. The interpolation error is $\mathcal{O}(h^{k+1})$ in the L^2-norm.
- A function u whose $(k+1)$-th derivative is identically zero can be interpolated exactly. These are all polynomial functions of degree $\leq k$ over Ω.
- The magnitude of the interpolation error does not depend on how large the derivatives of order $\leq k$ are. Only the $(k+1)$-th derivative matters.

Combining (3.21) with (3.20) we arrive at the estimate that we were looking for.

Corollary 3.1. *Under the hypotheses of Theorems 3.2 and 3.3, when h is small enough the finite element solution u_h built with P_k-elements satisfies, for some constant C depending on k but not on u or h,*

$$\|u - u_h\|_1 \leq C \, h^k \, \|u^{(k+1)}\|_0 \,. \tag{3.23}$$

Notice that the function and its k first derivatives need to be continuous for the theorem to apply, and the $(k+1)$-th derivative needs to be square-integrable.

We have thus proved that, using P_k-elements, u_h indeed converges to u in the H^1-norm and that the order of convergence is $\mathcal{O}(h^k)$.

We conclude this section with a partial proof of Theorem 3.3.

Proof. For brevity, we restrict the proof to the case $k = 1$. Consider a one-dimensional finite element mesh of P_1-elements over a domain $\Omega = [0, L]$, and a smooth function $u\colon [0, L] \to \mathbb{R}$. The mesh has n_{el} elements with domain $\Omega_e = [x_e, x_{e+1}]$ and

size $h_e = x_{e+1} - x_e > 0$ for $e = 1, \ldots, n_{\text{el}}$, with $x_1 = 0$ and $x_{n_{\text{el}}+1} = L$. The number of nodes is $m = n_{\text{el}} + 1$.

Let $\{N_1, \ldots, N_m\}$ be the hat functions over this mesh and $W_h = \text{span}\{N_1, N_2, \ldots, N_m\}$ the corresponding finite element space.

Definition 3.2. *We define the* **Lagrange finite element interpolant** *(or* **Lagrange interpolant** *for short)* $\mathscr{I}u$ *of* u *as*

$$\mathscr{I}u = \sum_{a=1}^{m} u(x_a) N_a. \tag{3.24}$$

More generally, the Lagrange interpolant $\mathscr{I}u$ *is the unique function of* $\mathscr{W}_h$ *that has as components the values of* u *at the nodes of the mesh.*

Notice that $\mathscr{I}u \in \mathscr{S}_h$, since $\mathscr{I}u \in \mathscr{W}_h$ and

$$\mathscr{I}u(0) = \sum_a u(x_a) N_a(0) = u(x_1) N_1(0) = u(0) N_1(0) = u(0) = g_0 \,.$$

We will estimate an upper bound for the *interpolation* errors $\|u - \mathscr{I}u\|_0$ and $\|u - \mathscr{I}u\|_1$ over $[0, L]$. To this end, we will split the errors as a sum of local errors over each element, namely,

$$\begin{aligned} \|u - \mathscr{I}u\|_0^2 &= \int_0^L |u - \mathscr{I}u|^2 \, dx \\ &= \sum_{e=1}^{n_{\text{el}}} \|u - \mathscr{I}u\|_{0,\Omega_e}^2 \end{aligned} \tag{3.25}$$

and

$$\begin{aligned} \|u - \mathscr{I}u\|_1^2 &= \int_0^L \left(|u - \mathscr{I}u|^2 + |u' - \mathscr{I}u'|^2 \right) \, dx \\ &= \|u - \mathscr{I}u\|_0^2 + \sum_{e=1}^{n_{\text{el}}} \|u' - \mathscr{I}u'\|_{0,\Omega_e}^2 \,. \end{aligned} \tag{3.26}$$

Let's see then how the error over an element can be obtained. Over element e consider the error function between u and $\mathscr{I}u$,

$$\eta(x) = u(x) - \mathscr{I}u(x) = u(x) - \left[u(x_e) N_1^e(x) + u(x_{e+1}) N_2^e(x) \right].$$

It satisfies that

(a) $\eta(x_e) = \eta(x_{e+1}) = 0$,

(b) for $x \in (x_e, x_{e+1})$ its derivative is

$$\eta'(x) = u'(x) - \mathscr{I}u'(x) = u'(x) - \frac{u(x_{e+1}) - u(x_e)}{h_e} \text{ and}$$

(c) its second derivative coincides with that of u, namely,

$$\eta''(x) = u''(x). \tag{3.27}$$

Since u is assumed smooth, because of (a) and of Rolle's theorem there exists $\xi \in (x_e, x_{e+1})$ such that $\eta'(\xi) = 0$. As a consequence,

$$\begin{aligned} |\eta'(x)| &= \left| \int_\xi^x \eta''(y)\, dy \right| \\ &= \left| \int_\xi^x u''(y)\, dy \right| \\ &\leq \left[\int_\xi^x u''(y)^2\, dy \right]^{\frac{1}{2}} \left[\int_\xi^x 1^2\, dy \right]^{\frac{1}{2}} \\ &\leq h_e^{\frac{1}{2}} \|u''\|_{0,\Omega_e} \end{aligned} \tag{3.28}$$

for all $x \in \Omega_e$. Notice now that

$$|\eta(x)| = \left| \int_{x_e}^x \eta'(y)\, dy \right| \leq h_e \max_{y \in \Omega_e} |\eta'(y)| \leq h_e^{\frac{3}{2}} \|u''\|_{0,\Omega_e}$$

which implies that

$$\|u - \mathcal{I}u\|_{0,\Omega_e} = \left[\int_{x_e}^{x_{e+1}} \eta(x)^2\, dx \right]^{\frac{1}{2}} \leq h_e^2 \|u''\|_{0,\Omega_e}\,.$$

Summing over the elements we get

$$\begin{aligned} \|u - \mathcal{I}u\|_0^2 = \sum_{e=1}^{n_{\text{el}}} \|u - \mathcal{I}u\|_{0,\Omega_e}^2 &\leq \sum_e h_e^4 \|u''\|_{0,\Omega_e}^2 \\ &\leq h^4 \sum_e \|u''\|_{0,\Omega_e}^2 = h^4 \|u''\|_0^2. \end{aligned} \tag{3.29}$$

and thus, since $\mathcal{I}u$ belongs to $\mathcal{W}_h$, (3.22) is proved with $C_I = 1$.

To prove (3.21) we use (3.28) to see that

$$\begin{aligned} \|u' - \mathcal{I}u'\|_{0,\Omega_e} &= \left[\int_{x_e}^{x_{e+1}} \eta'(x)^2\, dx \right]^{\frac{1}{2}} \\ &\leq h_e^{\frac{1}{2}} \|u''\|_{0,\Omega_e} \left[\int_{x_e}^{x_{e+1}} 1^2\, dx \right]^{\frac{1}{2}} \\ &= h_e \|u''\|_{0,\Omega_e}\,. \end{aligned}$$

Then,

$$\sum_{e=1}^{n_{\text{el}}} \|u' - \mathcal{I}u'\|_{0,\Omega_e}^2 \leq \sum_e h_e^2 \|u''\|_{0,\Omega_e}^2 \leq \left(\max_e h_e^2 \right) \sum_e \|u''\|_{0,\Omega_e}^2 = h^2 \|u''\|_0^2\,. \tag{3.30}$$

Inserting (3.30) and (3.29) into (3.26) we get

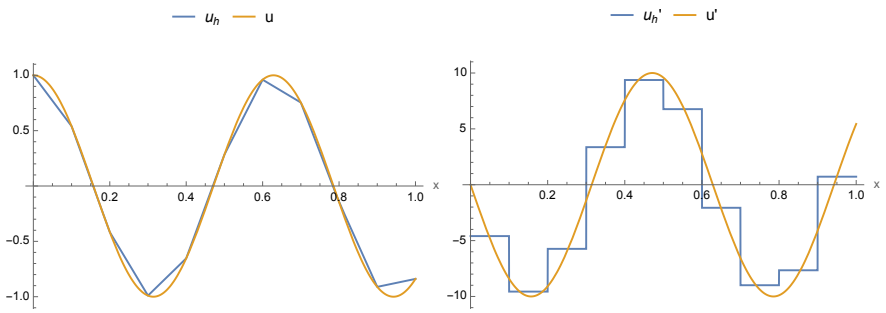
$$\|u - \mathcal{I}u\|_1 \leq \left(h^4 + h^2\right)^{\frac{1}{2}} \|u''\|_0\,.$$

At this point we must only assume that the mesh is fine enough, more specifically that $h \leq 1$ so that $h^4 \leq h^2$ and thus $\left(h^4 + h^2\right)^{\frac{1}{2}} \leq \sqrt{2}\, h^2$, to finalize the proof of (3.21) with $C_I = \sqrt{2}$.

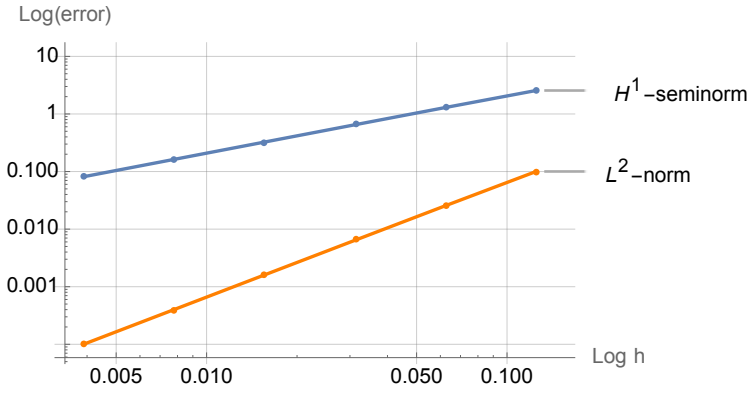
Since we have obtained two possible values (1 and $\sqrt{2}$) for C_I, we conclude that the theorem holds with the greater value, i.e., $C_I = \sqrt{2}$.

□

Example 3.1 Consider a mesh of $n_{\text{el}} \in \mathbb{N}$ equal-length P^1-elements over a domain $\Omega = [0,1]$, with $h = 1/n_{\text{el}}$, and $u(x) = \cos(10x) \in C^2([0,1])$. The interpolant and its first derivative are plotted below for $n_{\text{el}} = 11$.



The L^2-norms of the error $u - \mathcal{I}u$ and its derivative $u' - Iu'$ (termed the H^1-seminorm) are plotted next, in a log-log scale. Why a log-log plot? Because if we compute the log on both sides of a relation of the form error$\sim Ch^r$, we get that $\log(\text{error}) \sim \log C + r \log h$, so log(error) decreases linearly with $\log h$, and the slope of the line is equal to the order r.



Notice that the L^2-error decreases 2 units in the vertical axis per unit in the horizontal axis, a slope equal to 2, in agreement with the expected order of convergence in (3.29). Similarly, the H^1-seminorm decreases 1 unit in the vertical axis per unit in the horizontal one, indicating a slope equal to

1, in agreement with the first order of convergence in (3.30). The square of the H^1-norm is the sum of the squares of the two curves, and when plotted, it essentially overlaps with the curve of the H^1-seminorm, so it was not plotted. A quick inspection reveals that at the largest value of h, the L^2-norm of the error is already more than 10 times smaller than the H^1-seminorm, so when the two squared values are added, the error on the value of the function contributes less than 1% to the sum. The H^1-error is dominated by the error in the derivative.

3.3.3 Consequences of the Convergence in the H^1-norm

Let us go back to the estimate proved in the previous section:

$$\|u - u_h\|_1 \le C\, h^k\, \|u^{(k+1)}\|_0 \,. \tag{3.31}$$

From it, we can derive several useful consequences:

1. For any $x \in \Omega$,

$$|u(x) - u_h(x)| \le \left| \int_0^x (u'(s) - u_h'(s))\, ds \right| \le \|u' - u_h'\|_0\, L^{\frac{1}{2}} \le C(u)\, h^k \,. \tag{3.32}$$

   In other words, u_h converges **uniformly** in Ω to u. At any point x, $u_h(x)$ converges to $u(x)$ with order *at least* equal to $\mathcal{O}(h^k)$. In fact, the order is in general greater than k. This tells us that $u_h(x)$ is a **convergent approximation** of $u(x)$.

2. Assume that the problem represents a one-dimensional elastic problem, such as the vertical loading of a column by its own weight. In this case we would have $c(x) = 0$, $g_0 = 0$, $f(x) = -\rho A(x)\, g$ and $d_L = 0$, and u would represent the vertical displacement of the section at height x. The symbol ρ denotes de density, $A(x)$ the cross-sectional area at x, and g the acceleration of gravity. The elastic energy of such bar is given by

$$U(u) = \frac{1}{2} a(u,u), \tag{3.33}$$

   and it is important that $U(u_h)$ converges to $U(u)$ as $h \to 0$. This is readily obtained using the symmetry of $a(\cdot,\cdot)$,

$$\begin{aligned}
2U(u) - 2U(u_h) &= a(u,u) - a(u_h,u_h) - a(u,u_h) + a(u_h,u) \\
&= a(u, u-u_h) + a(u_h, u-u_h) \\
&= a(u+u_h, u-u_h) \\
&\le M \|u-u_h\|_1 \|u+u_h\|_1 \\
&\le M \|u-u_h\|_1 (2\|u\|_1 + \|u-u_h\|_1) \\
&\le 2MC\, h^k\, \|u^{(k+1)}\|_0 \|u\|_1 + \mathcal{O}(h^{2k}),
\end{aligned}$$

   which indeed guarantees convergence of the energy with order $\mathcal{O}(h^k)$.

3. Any integral of the form

$$I(u) = \int_\Omega g(u(x), u'(x))\, dx\,, \tag{3.34}$$

where g is twice continuously differentiable in its two variables, is approximated by $I(u_h)$ with order at least k but in general higher, as high as $\mathcal{O}(h^{2k})$.

Consider the example above of a column under its own weight. The change in gravitational energy of the column due to its deformation $u(x)$ is given by

$$G(u) = \int_\Omega \rho\, g\, A(x)\, u(x)\, dx\,. \tag{3.35}$$

This quantity is approximated by $G(u_h)$, in fact

$$G(u) - G(u_h) = \mathcal{O}(h^{2k})\,.$$

4. The convergence of $\|u - u_h\|_0$ to zero is also immediate from (3.31), since

$$\|u - u_h\|_0 \le \|u - u_h\|_1 \le C\, h^k\, \|u^{(k+1)}\|_0\,.$$

Notice that this proves convergence with order $\mathcal{O}(h^k)$. However, we know from (3.22) that it is possible to approximate u in the L^2-norm with order $\mathcal{O}(h^{k+1})$. The estimation above is thus **suboptimal**. But it turns out to be just a flaw of our demonstration. Using an elegant argument known as **Aubin-Nitsche trick** it can be proved that

$$\|u - u_h\|_0 \le C\, h^{k+1}\, \|u^{(k+1)}\|_0\,. \tag{3.36}$$

The finite element solution u_h converges with optimal order in the L^2-norm.

3.3.4 An Example of Numerical Convergence

Let us use the code developed in Chapter 1 to solve a particular case of Problem 1.1 for which we know the exact solution. This will allow us to confirm the predictions of our mathematical estimates.

A popular way to assess convergence is by constructing a **manufactured solution**. Here we will select the domain $[0,1]$, the coefficients

$$k(x) = 1, \qquad b(x) = 0, \qquad c(x) = 1$$

and the exact solution

$$u(x) = \sin(\alpha\, x^2)\,.$$

The trick is simply to set the source term $f(x)$ and the boundary conditions g_0 and d_L so that u solves the problem.

Noticing that

$$(ku')' = 2\alpha\left(\cos(\alpha x^2) - 2\alpha x^2 \sin(\alpha x^2)\right)$$

we compute that to satisfy the differential equation we need to set

$$f(x) = -2\alpha\left(\cos(\alpha x^2) - 2\alpha x^2 \sin(\alpha x^2)\right) + \sin(\alpha x^2),$$

$g_0 = u(x=0) = 0$, and

$$d_L = u'(x=L) = 2\alpha x \cos(\alpha x^2)\big|_{x=1} = 2\alpha\cos(\alpha)\,.$$

We select $\alpha = 20$.

With these data, we can readily use our code. Let us begin with a uniform mesh of 40 P_1 Lagrange elements. In Figure 3.1 we show the exact solution u, the finite element solution u_h and the P_1 Lagrange interpolant $\mathscr{I}u$, together with their derivatives. The interpolant $\mathscr{I}u$ coincides with u at the nodes. In part (c) of the figure we can see the interpolation error $u - \mathscr{I}u$, which has a parabola-like shape inside each element and is larger at elements with larger $|u''|$.

The approximation error $u - u_h$ (also shown in Figure 3.1(c)) can be seen as the sum of the interpolation error $u - \mathscr{I}u$ plus a numerical error $\mathscr{I}u - u_h$ that belongs to $\mathcal{V}_h$. Céa's lemma (Thm. 3.1) tells us something far from obvious: That the numerical error is bounded, up to a constant, by the interpolation error. In this case, the error $u' - u_h'$ is very close to $u' - (\mathscr{I}u)'$, so that the approximation error comes mostly from interpolation.

The errors in u and u', for a mesh with twice the number of elements, are shown in Figure 3.2. Notice from the vertical scales that the error in the function roughly gets divided by four with respect to the 40-element case, and the error in the derivative gets divided by two.

The convergence of u_h and $\mathscr{I}u$ towards u as h tends to zero, in L^2 and H^1 norms, can be seen in the following table:

Quantity	$h = 1/40$	$h = 1/80$	$h = 1/160$	$h = 1/320$	Order
$\|u - u_h\|_0$	0.3429	0.08327	0.02067	5.158E-3	$\mathcal{O}(h^2)$
$\|u - \mathscr{I}u\|_0$	0.0277	7.022E-3	1.762E-3	4.409E-4	$\mathcal{O}(h^2)$
$\|u' - u_h'\|_0$	3.6159	1.7920	0.8947	0.4477	$\mathcal{O}(h)$
$\|u' - \mathscr{I}u'\|_0$	3.5174	1.7805	0.8932	0.4475	$\mathcal{O}(h)$

As expected, many other quantities also converge as well. Here are some examples:

Quantity	$h = 1/40$	$h = 1/80$	$h = 1/160$	$h = 1/320$	Order
$u(L) - u_h(L)$	0.6720	0.1628	0.04038	0.01008	$\mathcal{O}(h^2)$
$u'(L) - u_h'(L)$	-11.469	-7.5888	-4.1838	-2.1783	$\mathcal{O}(h)$
$u'(L) - \mathscr{I}u'(L)$	-13.706	-8.0704	-4.2939	-2.2045	$\mathcal{O}(h)$
$\int(u - u_h)\,dx$	0.2909	0.07064	0.01752	4.373E-3	$\mathcal{O}(h^2)$
$\int(u - \mathscr{I}u)\,dx$	-8.670E-4	-2.135E-4	-5.318E-5	-1.327E-5	$\mathcal{O}(h^2)$

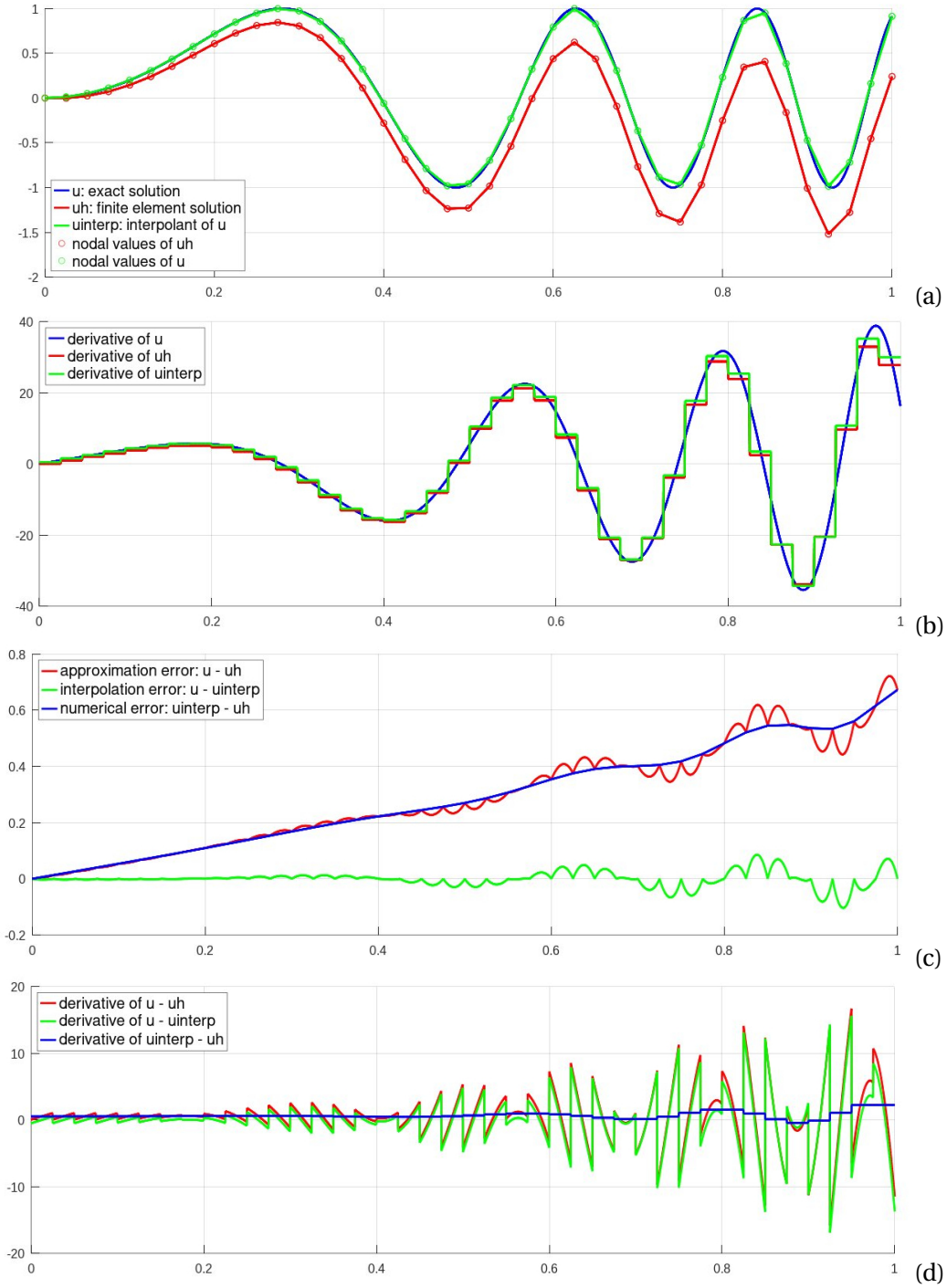


Figure 3.1 Example of numerical convergence. Results on a mesh of 40 elements, all of size $h = 1/40$. (a) Exact solution u, finite element solution u_h and interpolant of exact solution $\mathscr{I}u$. (b) Derivatives of u, u_h and $\mathscr{I}u$. (c) Approximation error $u - u_h$ and interpolation error $u - \mathscr{I}u$. (d) Errors in the derivative.

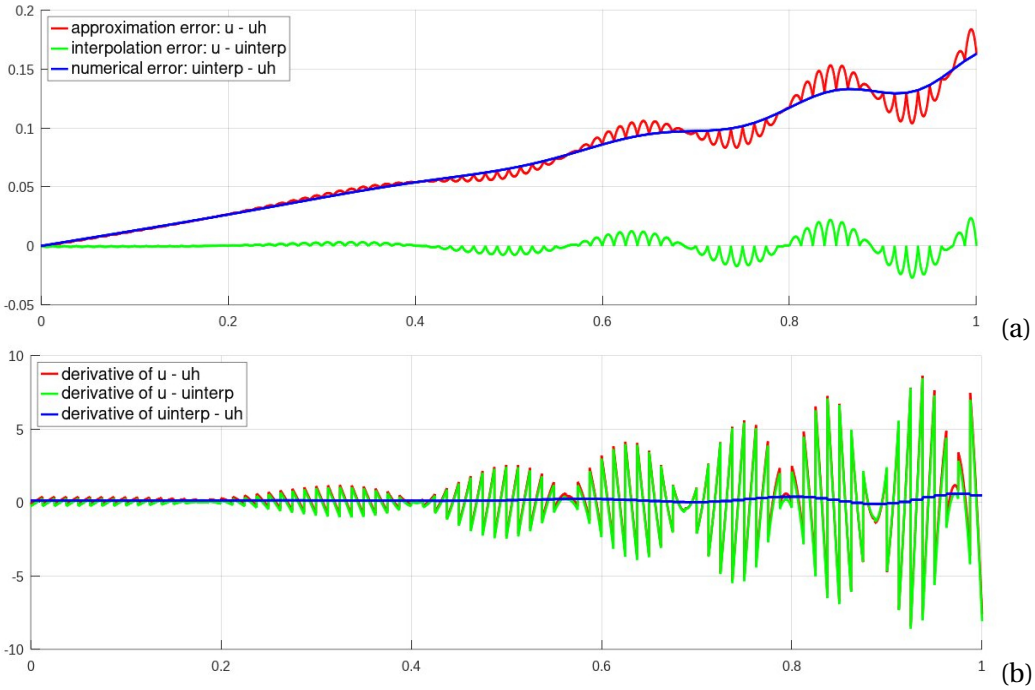


Figure 3.2 Example of numerical convergence. Results on a mesh of 80 elements, all of size $h = 1/80$. (a) Approximation error $u - u_h$ and interpolation error $u - \mathscr{I}u$. (b) Errors in the derivative. Notice the change in the vertical scales with respect to the previous figure.

All of these experimental results are consistent with the theory. Notice that $u_h'(L)$, for which we have proved no theoretical estimate, converges with order $\mathcal{O}(h)$. The energy $U(u) = \frac{1}{2}a(u,u)$ in this case equals $\frac{1}{2}(\|u'\|_0^2 + \|u\|_0^2)$ and converges with order $\mathcal{O}(h^2)$.

3.4 Fourth-order problems in one dimension

Finite element methods for fourth-order problems can also be analyzed using Theorem 3.1. This will explain, in particular, the need for continuously differentiable basis functions and thus justify the introduction of the Hermite finite element space.

To make things precise, let $u : [0, L] \to \mathbb{R}$ satisfy $u(0) = g_0$, $u'(0) = d_0$, $q(L)u'''(L) + q'(L)u''(L) = -W$ (W: applied load), $q(L)u''(L) = T$ (T: applied torque), together with

$$\big(q(x)u''(x)\big)'' + c(x)u(x) = f(x), \qquad \forall x \in \Omega\,. \tag{3.37}$$

The coefficient $q(x)$ must be greater than some $q_{\min} > 0$, while $c(x) \geq 0$. For the distributed load $f(x)$ we assume it to be square-integrable again.

We have already seen, in Problem 1.6, that appropriate bilinear and linear forms for this problem are

$$a(u,v) = \int_0^L \big(qu"v" + cuv\big)\,dx, \qquad \ell(v) = \int_0^L fv\,dx + W\,v(L) + T\,v'(L)\,.$$

The essential boundary conditions are those imposing $u(0)$ and $u'(0)$.

Let us now consider some finite element space $\mathcal{W}_h$ from which we define the trial and test spaces as

$$\mathcal{S}_h = \{w_h \in \mathcal{W}_h \,|\, w_h(0) = g_0,\, w_h'(0) = d_0\}, \qquad \mathcal{V}_h = \{w_h \in \mathcal{W}_h \,|\, w_h(0) = 0,\, w_h'(0) = 0\}.$$

The finite element solution $u_h \in \mathcal{S}_h$ is computed from

$$a(u_h, v_h) = \ell(v_h), \qquad \forall v_h \in \mathcal{V}_h\,. \tag{3.38}$$

We now retrace the steps followed in Sections 3.3.1-3.3.4 to analyze the convergence of u_h towards the exact solution u.

Exact consistency. The residual is given by

$$r(u, v_h) = \int_0^L \big(qu"v"_h + cuv_h\big)\,dx - \int_0^L fv_h\,dx - Wv_h(L) - Tv_h'(L)$$

which upon integrating twice by parts in each element transforms into

$$\begin{aligned} r(u,v_h) \;=\; & \int_0^L \big((qu")" + cu - f\big)\, v_h\, dx \\ & - \big(q'(L)u''(L) + q(L)u'''(L) + W\big)\, v_h(L) \\ & + \big(q(L)u''(L) - T\big)\, v_h'(L) \\ & + \big(q'(0)u''(0) + q(0)u'''(0)\big)\, v_h(0) \\ & - q(0)u''(0)\, v_h'(0) \\ & - \sum_z \big(q'(z)u''(z) + q(z)u'''(z)\big)\big(v_h(z^-) - v_h(z^+)\big) \\ & + \sum_z q(z)u''(z)\big(v_h'(z^-) - v_h'(z^+)\big)\ , \end{aligned}$$

with z again running over all interelement boundaries. The first, second and third lines of the right-hand side above are zero because u satisfies the differential equation and the (natural) boundary conditions at L. For the fourth and fifth lines to be zero one needs to impose $v_h(0) = v_h'(0) = 0$, which justifies the definition of $\mathcal{V}_h$ above. For the sixth line to be zero, the function v_h must be continuous at z (zero jump in the function). Finally, for the last line to be zero, **the derivative v_h' must be continuous at** z (zero jumps in the derivative across interelement boundaries). This is a generalization of the result we obtained after evaluating consistency in §1.5.3.2.

As a consequence, for consistency **the finite element space $\mathcal{W}_h$ must consist of functions that are continuous and have continuous derivative**. This is why the simplest space for this problem is the Hermite H_3 piecewise cubic finite element space.

Coercivity. The bilinear form $a(\cdot,\cdot)$ turns out to be coercive in the so-called H^2-norm, which is given by

$$\|v_h\|_2 = \left[\int_0^L \big(v_h''(x)^2 + v_h'(x)^2 + v(x)^2\big)\, dx\right]^{\frac{1}{2}} = \big(\|v_h\|_0^2 + \|v_h'\|_0^2 + \|v_h''\|_0^2\big)^{\frac{1}{2}}\ . \quad (3.39)$$

This requires that $q_{\min} > 0$, in agreement with physical constraints.

Continuity. The continuity of $a(\cdot,\cdot)$ and $\ell(\cdot)$ in the H^2-norm also holds, and can be proved with the same arguments used for the second order case. Naturally, some requirements appear in the coefficients: q and c must be bounded and the distributed load f must have a finite integral. These requirements are physically sound and hold in most cases.

All hypothesis of Thm. 3.1 have been checked, from which we conclude:

Theorem 3.4. *If the finite element space $\mathcal{W}_h$ consists of* **continuously differentiable functions** *that in addition are C^2 in each element, then the finite element solution u_h defined by* (3.38) *exists, is unique, and satisfies*

$$\|u - u_h\|_2 \le C \min_{w_h \in \mathcal{S}_h} \|u - w_h\|_2 \quad (3.40)$$

with the constant C independent of the adopted mesh.

We now take as $\mathcal{W}_h$ the **Hermite** H_3 **finite element space** introduced in section 1.5.4, which satisfies the hypotheses of the theorem since it is a piecewise cubic polynomial space contained in $C^1([0,L])$.

There is a corresponding **Hermite interpolant** of the exact solution u (assumed smooth), given by

$$\mathcal{I}u(x) = u(x_1)H_1(x) + u'(x_1)H_2(x) + u(x_2)H_3(x) + u'(x_2)H_4(x) + \ldots \tag{3.41}$$

Using this interpolant and following analogous steps to those in the proof of Theorem 3.3 one arrives at

Theorem 3.5. *Let u be a C^4 function in $\Omega = [0,L]$ such that $u(0) = g_0$ and $u'(0) = d_0$. Let $\mathcal{W}_h$ be the cubic Hermite finite element space built on a mesh $\mathcal{T}_h$ of Ω. Let h denote the length of the largest element in $\mathcal{T}_h$. Then, there exists a constant C_I such that*

$$E_2(\mathcal{S}_h, u) = \min_{w_h \in \mathcal{S}_h} \|u - w_h\|_2 \le C_I\, h^2\, \|u^{(4)}\|_0\,, \tag{3.42}$$

$$E_1(\mathcal{S}_h, u) = \min_{w_h \in \mathcal{S}_h} \|u - w_h\|_1 \le C_I\, h^3\, \|u^{(4)}\|_0\,, \tag{3.43}$$

$$E_0(\mathcal{S}_h, u) = \min_{w_h \in \mathcal{S}_h} \|u - w_h\|_0 \le C_I\, h^4\, \|u^{(4)}\|_0\,, \tag{3.44}$$

where $u^{(4)} = d^4u/dx^4$.

Combining Theorems 3.4 and 3.5 we conclude that the finite element solution u_h satisfies, for some constant $C > 0$,

$$\|u - u_h\|_2 \le C\, h^2\, \|u^{(4)}\|_0\,, \tag{3.45}$$

and thus converges toward u as the mesh is refined. The order of convergence is $\mathcal{O}(h^2)$ in the H^2-norm. It is possible to prove that the convergence is also optimal in the H^1- and L^2-norms, with orders $\mathcal{O}(h^3)$ and $\mathcal{O}(h^4)$, respectively.

All the consequences of convergence in the H^1-norm thus hold, in particular that u_h converges **uniformly** to u in Ω, with order at least $\mathcal{O}(h^3)$ but in fact higher.

Additionally, we have that u_h' converges **uniformly** to u' in Ω, with order at least $\mathcal{O}(h^2)$. So, in fourth-order problems, $u_h'(x)$ approximates $u'(x)$ at all points and uniformly.

To confirm these statements numerically, let us again apply the method of manufactured solutions. We select the constants $q = 1$, $c = c_0$ and specify the solution

$$u(x) = \sin(\alpha x^2)$$

in the domain $[0,L]$ (with $L = 1$) so that $g_0 = u(0) = 0$ and $d_0 = u'(0) = 0$. By differentiating u we compute the appropriate distributed load f, end load W and end torque T.

$$f = u^{(4)} + c_0 u = 4\alpha^2\left((4\alpha^2 x^4 - 3)\sin(\alpha x^2) - 12\alpha x^2\cos(\alpha x^2)\right) + c_0 \sin(\alpha x^2)\,,$$

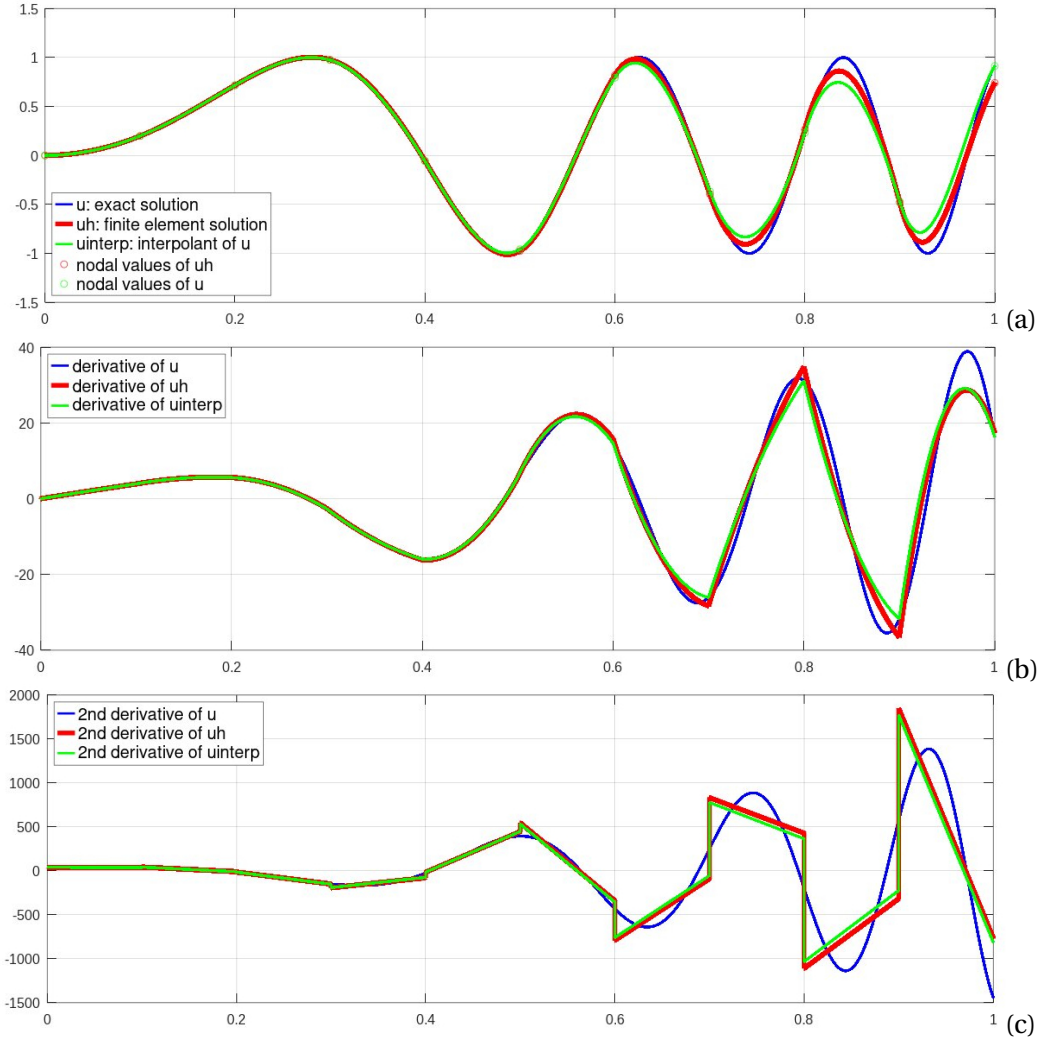


Figure 3.3 Example of numerical convergence. Results on a mesh of 10 Hermite cubic elements, all of size $h = 1/10$. (a) Exact solution u, finite element solution u_h and interpolant of exact solution $\mathscr{I}u$. (b) First derivative of u, u_h and $\mathscr{I}u$. (c) Second derivative of u, u_h and $\mathscr{I}u$.

$$W = -u^{(3)}(1) = 4\alpha^2\,(3\sin\alpha + 2\alpha\cos\alpha)\ ,$$

$$T = u''(1) = 2\alpha\,(\cos\alpha - 2\alpha\sin\alpha)\ .$$

We consider $c_0 = 10^6$ and $\alpha = 20$. We then run the code developed in Section 1.5, slightly modified so as to consider f not constant within each element. We begin with a uniform mesh of 20 Hermite cubic elements, so that $h = 1/20$. In Figure 3.3 we show the exact solution u, the finite element solution u_h and the H_3 interpolant $\mathscr{I}u$, together with their derivatives. The corresponding errors are shown in Fig. 3.4.

The convergence of u_h and $\mathscr{I}u$ towards u as h tends to zero, in L^2, H^1 and

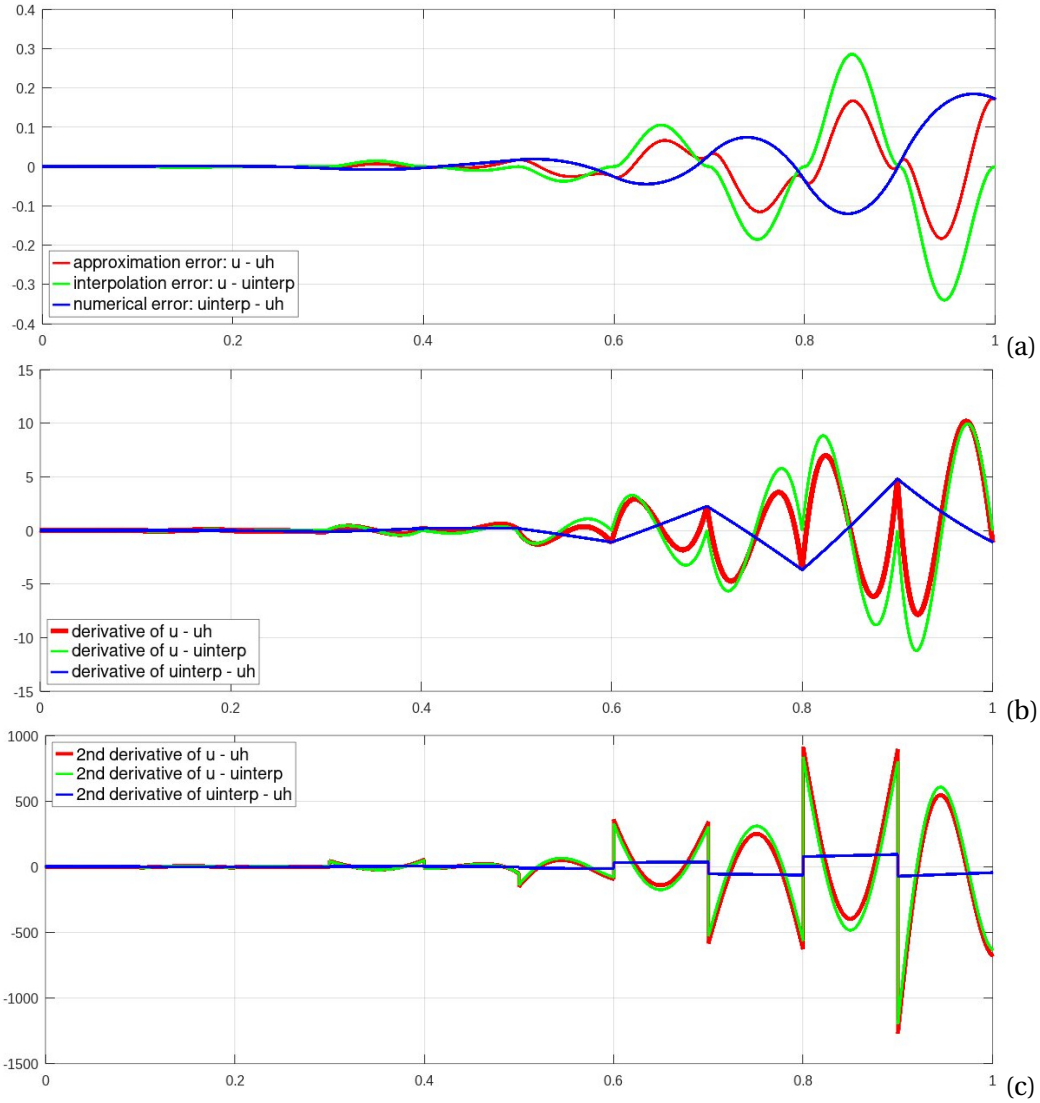


Figure 3.4 Example of numerical convergence. Results on a mesh of 10 Hermite cubic elements, all of size $h = 1/10$. (a) Approximation error $u - u_h$, interpolation error $u - \mathscr{I}u$ and numberical error $\mathscr{I}u - u_h$. (b) First derivative of the errors in (a). (c) Second derivative of the errors in (a).

H^2 norms, can be seen in the following table:

Quantity	$h = 1/10$	$h = 1/20$	$h = 1/40$	$h = 1/80$	$h = 1/160$	Order
$\|u - u_h\|_0$	0.0553	4.029E-3	2.750E-4	1.734E-5	1.086E-6	$\mathcal{O}(h^4)$
$\|u - \mathcal{I}u\|_0$	0.0983	5.302E-3	3.688E-4	2.350E-5	1.476E-6	$\mathcal{O}(h^4)$
$\|u' - u_h'\|_0$	2.7243	0.3536	0.0506	6.499E-3	8.175E-4	$\mathcal{O}(h^3)$
$\|u' - \mathcal{I}u'\|_0$	3.4610	0.3707	0.0512	6.517E-3	8.180E-4	$\mathcal{O}(h^3)$
$\|u'' - u_h''\|_0$	225.92	48.424	13.291	3.3800	0.8484	$\mathcal{O}(h^2)$
$\|u'' - \mathcal{I}u''\|_0$	222.44	48.385	13.290	3.3800	0.8484	$\mathcal{O}(h^2)$

Numerical convergence in other quantities is as follows:

Quantity	$h = 1/10$	$h = 1/20$	$h = 1/40$	$h = 1/80$	$h = 1/160$	Order
$u(L) - u_h(L)$	0.1726	-2.198E-4	-2.661E-4	-2.007E-5	-1.305E-6	$\mathcal{O}(h^4)$
$u'(L) - u_h'(L)$	-1.0843	-0.1365	-0.0158	-1.098E-3	-7.026E-5	$\mathcal{O}(h^4)$?
$u''(L) - u_h''(L)$	-676.63	103.00	72.175	23.577	6.5204	$\mathcal{O}(h^2)$
$u''(L) - \mathcal{I}u''(L)$	-631.51	102.25	72.137	23.576	6.5204	$\mathcal{O}(h^2)$
$\int(u - u_h)\,dx$	-3.037E-6	-3.461E-8	-1.061E-9	4.528E-11	-8.328E-12	$\mathcal{O}(h^5)$?
$\int(u - \mathcal{I}u)\,dx$	-8.554E-3	-3.034E-4	-1.709E-5	-1.042E-6	-6.476E-8	$\mathcal{O}(h^4)$

These experimental results are consistent with the theory. Though for the second derivative $u_h''(L)$ we have no theoretical estimate, it converges with order $\mathcal{O}(h^2)$. The energy $U(u) = \frac{1}{2}a(u,u)$ equals $\frac{1}{2}(\|u''\|_0^2 + c_0\|u\|_0^2)$ and converges with order $\mathcal{O}(h^4)$. It is interesting to notice the error of $\int u_h\,dx$, which is much smaller than that of $\int \mathcal{I}u\,dx$. The integral of the numerical solution is superconvergent (it converges faster than the interpolation estimate).

3.5 Second-Order Problems in Two Dimensions

As you have seen in the previous sections, the strategy for analyzing a finite element method consists of:
(a) Checking the hypotheses of Céa's lemma (Theorem 3.1), followed by
(b) Applying an interpolation result to estimate the convergence of $\min_{w_h \in \mathcal{S}_h} \|u - w_h\|$. The norm to be used in (b) is determined while performing (a).

We will now pursue this strategy for second order problems in 2D, for which a finite element method was defined in Chapter 2. Let Ω be a two-dimensional

domain and let u be a smooth function satisfying

$$-\operatorname{div}(K\nabla u) + c\,u = f\,, \qquad \text{in } \Omega, \tag{3.46}$$

together with

$$u = g\,, \qquad \text{on } \partial\Omega_D, \text{ and} \tag{3.47}$$

$$(K\nabla u)\cdot \check{n} = H \qquad \text{on } \partial\Omega_N\,. \tag{3.48}$$

We assume, as in Problem 2.1 that $\partial\Omega_D$ and $\partial\Omega_N$ are disjoint and that their union covers $\partial\Omega$. The diffusion matrix $K(x)$ is assumed bounded, symmetric and positive definite, with all eigenvalues greater than some $k_{\min} > 0$. The coefficient $c(x)$ is assumed bounded and non-negative.

The finite element method introduced in 2.4 reads, as usual: *Find $u_h \in \mathcal{S}_h$ such that*

$$a(u_h, v_h) = \ell(v_h) \tag{3.49}$$

for all $v_h \in \mathcal{V}_h$, where

$$a(u_h, v_h) = \int_\Omega (K\nabla u_h\cdot\nabla v_h + c\,u_h\,v_h)\; d\Omega\,, \tag{3.50}$$

$$\ell(v_h) = \int_\Omega f\,v_h\,d\Omega + \int_{\partial\Omega_N} H\,v_h\,d\Gamma\,, \tag{3.51}$$

and

$$\mathcal{S}_h = \{w_h \in \mathcal{W}_h \mid w_h = g \text{ on } \partial\Omega_D\}\,, \tag{3.52}$$

$$\mathcal{V}_h = \{w_h \in \mathcal{W}_h \mid w_h = 0 \text{ on } \partial\Omega_D\}\,. \tag{3.53}$$

3.5.1 Checking the Hypotheses of Céa's Lemma

Exact consistency. The consistency residual is

$$\begin{aligned} r(u, v_h) &= a(u, v_h) - \ell(v_h) \\ &= \int_\Omega \left(K\nabla u\cdot\nabla v_h + cuv_h - fv_h\right)\, d\Omega - \int_{\partial\Omega_N} Hv_h\,d\Gamma\,. \end{aligned}$$

We know that u is smooth by hypotheses and that v_h, belonging to the finite element space $\mathcal{W}_h$, is smooth inside each element. We want to determine what the required continuity at interelement boundaries is, and thus we assume no a priori continuity. We can nonetheless integrate by parts **element-wise**, which results in

$$\begin{aligned} r(u, v_h) = &\sum_e \int_{\Omega_e} \left(-\operatorname{div}K\nabla u + cu - f\right)\, v_h\, d\Omega \\ &+ \int_{\partial\Omega_N} (K\nabla u\cdot\check{n} - H)\; v_h\, d\Gamma \\ &+ \int_{\partial\Omega_D} (K\nabla u\cdot\check{n})\; v_h\, d\Gamma \\ &+ \sum_a \int_{\gamma_a} [\![v_h]\!] K\nabla u\cdot\check{n}\; d\gamma\,. \end{aligned} \tag{3.54}$$

The first and second terms are automatically zero because u is an exact solution. The third term is zero because v_h is zero on $\partial\Omega_D$.

The fourth term is a sum over all inter-element boundaries (edges), where $[\![v_h]\!]$ is the jump in v_h across the edge γ_a. The properties of u do not guarantee at all that this term vanishes, and thus **for exact consistency to hold, the functions in $\mathcal{V}_h$ must be continuous at inter-element boundaries**.

Coercivity. Being a second order elliptic problem, the norm in which coercivity can be established is the H^1-norm, which reads

$$\|v_h\|_1 = \left[\int_\Omega \left(\|\nabla v_h(x)\|^2 + v_h(x)^2\right)\, d\Omega\right]^{\frac{1}{2}} . \tag{3.55}$$

The notation $\|\nabla v_h(x)\|$ refers to the euclidean norm of the vector $\nabla v_h(x)$. It is frequent to consider the L^2-norm of a vector field $w : \Omega \to \mathbb{R}^d$, which is defined as

$$\|w\|_0 = \left[\int_\Omega \|w(x)\|^2\, d\Omega\right]^{\frac{1}{2}} . \tag{3.56}$$

With this notation, it holds that $\|v_h\|_1^2 = \|\nabla v_h\|_0^2 + \|v_h\|_0^2$.

If $c(x) \geq c_{\min} > 0$ it is immediate to prove that

$$a(v_h, v_h) = \int_\Omega \left(K\nabla v_h \cdot \nabla v_h + c v_h^2\right) \geq \min(k_{\min}, c_{\min})\, \|v_h\|_1^2$$

for all $v_h \in \mathcal{V}_h$. Thus $a(\cdot,\cdot)$ is coercive wich coercivity constant $\alpha = \min(k_{\min}, c_{\min})$.

In pure diffusion problems (i.e., problems with $c(x) = 0$ for all x) the coercivity of $a(\cdot,\cdot)$ is less evident and depends on the boundary conditions. For example, if there is no Dirichlet boundary ($\partial\Omega_D = \emptyset$), then constant functions belong to $\mathcal{V}_h$. Let $v_h(x) = A \in \mathbb{R}$ for all $x \in \Omega$ be a constant function, with $A \neq 0$. Then, when $c(x) = 0$ for $x \in \Omega$,

$$a(v_h, v_h) = \int_\Omega k\|\nabla v_h\|^2\, d\Omega = 0$$

and

$$\|v_h\|_1^2 = \int_\Omega A^2\, d\Omega > 0\,,$$

contradicting coercivity.

This shows that, if $c(x) \equiv 0$, boundary conditions must be such that **constant functions do not belong to $\mathcal{V}_h$**. Such a condition is generally met because in purely diffusive problems the **problem is ill-posed if the measure (length) of $\partial\Omega_D$ is zero**. If $\partial\Omega_D$ has positive length (and the domain is bounded), then the so-called **Poincaré inequality** guarantees that there exists $C_P > 0$ (independent of h) such that

$$\|v_h\|_0^2 \leq C_P \int_\Omega \|\nabla v_h\|^2\, d\Omega\,. \tag{3.57}$$

Using (3.57) it is easy to prove that $a(\cdot,\cdot)$ is coercive with just the hypothesis $c_{\min} \geq 0$. In fact,

$$\begin{aligned} a(v_h, v_h) &= \int_\Omega \left(K \nabla v_h \cdot \nabla v_h + c v_h^2 \right) \\ &\geq k_{\min} \int_\Omega \|\nabla v_h\|^2 \, d\Omega + c_{\min} \|v_h\|_0^2 \\ &\geq \frac{k_{\min}}{2} \int_\Omega \|\nabla v_h\|^2 \, d\Omega + c_{\min} \|v_h\|_0^2 + \frac{k_{\min}}{2C_P} \|v_h\|_0^2 \\ &\geq \min\left(\frac{k_{\min}}{2}, c_{\min} + \frac{k_{\min}}{2C_P} \right) \|v_h\|_1^2 \, . \end{aligned}$$

We see that there is coercivity, with the coercivity constant independent of the mesh and not depending on $c_{\min} > 0$.

Remark: To preclude non-zero constant functions from belonging to $\mathcal{V}_h$ it would suffice to fix just one nodal value of v_h to zero (that is, to have $\partial\Omega_D$ equal to a single point). In such a case the bilinear form is indeed coercive in $\mathcal{V}_h$, but the coercivity constant tends to zero as $h \to 0$. This reflects a mathematical fact about the exact problem, which is not well-posed if the length of $\partial\Omega_D$ is zero.

Continuity: The continuity of $a(\cdot,\cdot)$ in $H^1(\Omega)$ can be proved in much the same way as done in the one-dimensional case. In fact, one obtains

$$|a(u - w_h, v_h)| \leq (k_{\max} + c_{\max}) \|u - w_h\|_1 \, \|v_h\|_1 \, ,$$

where

$$k_{\max} = \max_{x \in \Omega} \max_{i,j} K_{ij}(x) \, .$$

What about the continuity of the linear functional? Remember that

$$\ell(v_h) = \int_\Omega f \, v_h \, d\Omega + \int_{\partial\Omega_N} H \, v_h \, d\Gamma \, ,$$

The first term is certainly continuous if $f \in L^2(\Omega)$, since by Cauchy-Schwartz inequality in $L^2(\Omega)$ we have that

$$\int_\Omega f \, v_h \, d\Omega \leq \|f\|_0 \, \|v_h\|_0 \leq \|f\|_0 \, \|v_h\|_1 \, .$$

If the source term function $f : \Omega \to \mathbb{R}$ is not square-integrable this term can still be continuous in $H^1(\Omega)$, but the proof is more technical.

The second term of $\ell(\cdot)$ involves a one-dimensional integral over the boundary segment (or segments) $\partial\Omega_N$. As all one-dimensional integrals, it satisfies Cauchy-Schwartz inequality, so

$$\int_{\partial\Omega_N} H \, v_h \, d\Gamma \leq \|H\|_{L^2(\partial\Omega_N)} \, \|v_h\|_{L^2(\partial\Omega_N)} \, .$$

The last ingredient needed in the proof is a so-called **trace theorem**. The version we will use here, that we accept without proof, reads: "If $v_h \in \mathcal{W}_h$, where $\mathcal{W}_h$ is a continuous finite element space (i.e., $\mathcal{W}_h \subset C^0(\Omega)$ and piecewise polynomial), then for any subset Γ of $\partial\Omega$ there exists $C_T > 0$ **independent of the mesh** such that

$$\|v_h\|_{L^2(\Gamma)} \leq C_T \|v_h\|_1 . " \tag{3.58}$$

Many trace inequalities exist. They all share the structure of (3.58), in that a norm of the function evaluated at the boundary of Ω is bounded by a norm that considers the function over the interior of Ω.

Collecting the previous results, we have proved that

$$\ell(v_h) \leq m \|v_h\|_1$$

for all $v_h \in \mathcal{W}_h$, with

$$m = \|f\|_0 \ + \ C_T \|H\|_{L^2(\partial\Omega_N)} . \tag{3.59}$$

3.5.2 Convergence

In the previous section all hypotheses of Céa's lemma have been checked with constants independent of the mesh (with some additional hypotheses that appeared along the way). We thus know that there exists C such that

$$\|u - u_h\|_1 \ \leq \ C \min_{w_h \in \mathcal{S}_h} \|u - w_h\|_1 . \tag{3.60}$$

Let us now consider a specific finite element space. We assume for now that the domain Ω is polygonal. We take as $\mathcal{W}_h$ the space of P_k **Lagrange finite elements** in two dimensions, of which the mesh $\mathcal{T}_h$ consists of triangles (and has no hanging nodes). This family of finite elements is very popular and general (any polygonal domain can be decomposed into triangles), and extends readily to dimensions greater than two.

Under a suitable set of hypotheses, it is possible to prove that, using P_k Lagrange elements,

$$E_1(\mathcal{S}_h, u) = \min_{w_h \in \mathcal{S}_h} \|u - w_h\|_1 \leq C_I \, h^k \, \|D^{k+1} u\|_0 , \tag{3.61}$$

and

$$E_0(\mathcal{S}_h, u) = \min_{w_h \in \mathcal{S}_h} \|u - w_h\|_0 \leq C_I \, h^{k+1} \, \|D^{k+1} u\|_0 , \tag{3.62}$$

for some $C_I > 0$ independent of u and of the mesh $\mathcal{T}_h$. In other words, the same estimates of Theorem 3.3 also hold for the two-dimensional case, with the only difference that the $(k+1)$-th derivative is now a tensor.

The necessary hypotheses, however, are much more technical than in 1D. Let us discuss them in some detail:

- **Regularity of the mesh.** The size h_e of each element is defined as the diameter of the subdomain Ω_e. The diameter of its inscribed circle is denoted by ρ_e. The mesh size is defined as $h = \max_e h_e$. A mesh is said to be **regular** if $\max_e h_e/\rho_e$ is bounded by some fixed constant. This essentially means that there are no small angles in the mesh, irrespective of how fine the mesh $\mathscr{T}_h$ is. **We adopt here the hypothesis that $\mathscr{T}_h$ is regular as $h \to 0$,** which is sufficient for E_1 to obey (3.61) if the other hypotheses are met.

  Remark: Regularity of the mesh is not strictly necessary for E_1 to satisfy (3.61) with P_k Lagrange elements. For P_1 elements it is known that a weaker condition than mesh regularity, called *maximum angle condition* (maximum angle $\leq \gamma < \pi$ for all meshes, irrespective of h), is sufficient for E_1 converge to zero with order h [1, 6].

- **Accurate approximation of the Dirichlet boundary condition.** Notice that, remembering that Ω is a polygon, the Dirichlet condition $u = g$ is imposed along straight segments. Let s be the arc-length tangential coordinate along one of the segments. If g, viewed as a function of s, is not continuous and piecewise polynomial, then there is no function w_h in $\mathscr{W}_h$ that satisfies $w_h = g$ on $\partial\Omega_D$. This would mean that $\mathscr{S}_h$ is empty!! For this reason the space $\mathscr{S}_h$ must be defined with an interpolation (or some sort of approximation) of g, that we will denote by g_h. The subscript h indicates that g_h in general depends on the mesh. The definition becomes

$$\mathscr{S}_h = \left\{ w_h \in \mathscr{W}_h \mid w_h = g_h \text{ on } \partial\Omega_D \right\} . \tag{3.63}$$

  For E_1 to be of order h^k it is **necessary** that g_h is close enough to g, more specifically that

$$\|g - g_h\|_{L^2(\partial\Omega_D)} \leq C_D\, h^k \tag{3.64}$$

  for some $C_D > 0$. Fortunately, if g_h is the P_k **Lagrange interpolant** of g (both viewed as functions of s) **and g is smooth enough**, then (3.64) is automatically satisfied.

  In fact, for the P_k Lagrange interpolant it holds that

  (a) $\|g - g_h\|_{L^2(\partial\Omega_D)}$ is of order h^{k+1} if $g''(s)$ is in $L^2(\partial\Omega_D)$. If g has no singularities, this requirement is equivalent to g being C^1.

  (b) $\|g - g_h\|_{L^2(\partial\Omega_D)}$ is of order h^k if $g'(s)$ is in $L^2(\partial\Omega_D)$. If g has no singularities, this requirement is equivalent to g being C^0.

  The proof can be easily adapted from that of (3.29), since $\partial\Omega_D$ is a one-dimensional manifold (possibly consisting of several disjoint parts).

With this we arrive at

Theorem 3.6. Convergence of P_k Lagrange finite element approximation for elliptic second order problems in 2D. *Let u be the solution to (3.46)-(3.48), supposed smooth enough for its $(k+1)$-th derivatives to be in $L^2(\Omega)$. Let the coefficients*

satisfy

$$0 < k_{\min} \le k(x) \le k_{\max} < +\infty\,, \qquad in\,\Omega\,, \tag{3.65}$$

$$0 \le c_{\min} \le c(x) \le c_{\max} < +\infty\,, \qquad in\,\Omega\,, \tag{3.66}$$

$$f \in L^2(\Omega)\,, \tag{3.67}$$

$$H \in L^2(\partial\Omega_N)\,. \tag{3.68}$$

Let u_h be the solution to (3.49)-(3.53), with $\mathcal{W}_h$ being the P_k **Lagrange finite element space** *associated with a* **regular triangulation** *$\mathcal{T}_h$.*

Assume that (3.52) has been replaced by (3.63), where g_h is the P_k **Lagrange interpolant** *of g along $\partial\Omega_D$.*

Then, there exists a constant $C(u) > 0$, dependent on u but independent of the mesh size h, such that

$$\|u - u_h\|_1 \le C(u)\, h^k\,. \tag{3.69}$$

Proof. **(a)** The proof when the interpolation error of the Dirichlet condition is zero, that is, when $g_h = g$, is immediate. Of course for this to happen the boundary data g must be a polynomial of degree $\le k$. The estimate (3.69) follows directly from (3.60) and (3.61).

(b) When $g \ne g_h$ the proof requires a little more work. Let us assume for simplicity that $\partial\Omega_D = \partial\Omega$. Let θ be the *exact solution* of (3.46)-(3.48) when $f = 0$ and the boundary data is $g - g_h$. We can decompose $u = u_h^* + \theta$, u_h^* defined as the *exact solution* corresponding to boundary data g_h. We know that $\|u_h^* - u_h\|_1 \le C\, h^k$ because of (a) above. A stability inequality from the theory of elliptic PDEs establishes that

$$\|\theta\|_1 \le C_2\, \|\theta\|_{L^2(\partial\Omega)} = C_2\, \|g - g_h\|_{L^2(\partial\Omega)}\,,$$

which combined with (3.64) yields $\|\theta\|_1 \le C_3 h^k$.

□

Thus, if the meshing algorithm is regular and the boundary conditions are regular and well approximated, the numerical solution u_h converges to the exact solution u in the sense of the H^1-norm.

Any quantity that is continuous in the H^1-norm will converge as well. For example, the H^1-norm itself will satisfy $\|u_h\|_1 \overset{h\to 0}{\to} \|u\|_1$.

3.5.3 Numerical example: The uniformly heated square rod with a hot lid

Consider u to be the solution of $-\Delta u = 1$ in the unit square domain $\Omega = (0,1) \times (0,1)$, with Dirichlet boundary conditions $u = 0$ over the left, bottom and right sides, and $u = \delta$ over the top side (the lid).

Under a thermal interpretation, u is the temperature field of a uniformly heated square bar with three of its sides in contact with a thermostat at temperature $u = 0$ and the remaining side in contact with a thermostat at $u = \delta$.

The problem seems physically possible, though perfect thermal contact with perfect thermostats are ideal boundary conditions that cannot be realized in practice.

We would like to predict its solution, approximately of course. For this purpose we build a sequence of finite element meshes M1, ..., M5, with mesh size approximately $h = 0.25, 0.25/2, \ldots, 0.25/16$. On each mesh we compute the finite element solution u_h defined by (3.49)-(3.53), with

$$K = 1, \;\; c = 0, \;\; f = 1, \;\; \partial\Omega_D = \partial\Omega$$

and the function g equal to δ on the upper side and zero elsewhere. We use the P_1 Lagrange finite element code of Chapter 2.

We will discuss two values of δ, namely $\delta = 0$ and $\delta = 0.08$. The corresponding numerical solutions for meshes $M1$, $M3$ and $M5$ are shown in Fig. 3.5.

By direct inspection of Fig. 3.5 we can argue that the numerical solutions seem to converge to some smooth function u which, if the method makes sense, must be the exact solution of the PDE.

Numerical solutions are computed so as to estimate some quantity of interest $Q(u)$ by its approximation $Q(u_h)$. The method is useful (for this quantity) if $\lim_{h\to 0} Q(u_h) = Q(u)$. For this to happen, it is first necessary that the sequence $Q(u_h)$ converges to something, and this is what we are going to check.

Let us consider the following quantities of interest:

$$Q_1(u_h) = \int_\Omega |u_h(x)|\, d\Omega, \tag{3.70}$$

$$Q_2(u_h) = \left[\int_\Omega |u_h(x)|^2\, d\Omega\right]^{\frac{1}{2}}, \tag{3.71}$$

$$Q_3(u_h) = \sup_{x\in\Omega} |u_h(x)|, \tag{3.72}$$

$$Q_4(u_h) = \int_\Omega \|\nabla u_h(x)\|\, d\Omega, \tag{3.73}$$

$$Q_5(u_h) = \left[\int_\Omega \|\nabla u_h(x)\|^2\, d\Omega\right]^{\frac{1}{2}}, \tag{3.74}$$

$$Q_6(u_h) = \sup_{x\in\Omega} \|\nabla u_h(x)\|. \tag{3.75}$$

Quantities $Q_1 - Q_3$ correspond to $\|u_h\|_{L^p(\Omega)}$, with $p = 1$, 2 and ∞. Quantities $Q_4 - Q_6$ are analogous for $\|\nabla u_h\|_{L^p(\Omega)}$.

The values obtained for the five meshes, in the case $\delta = 0$, were the following:

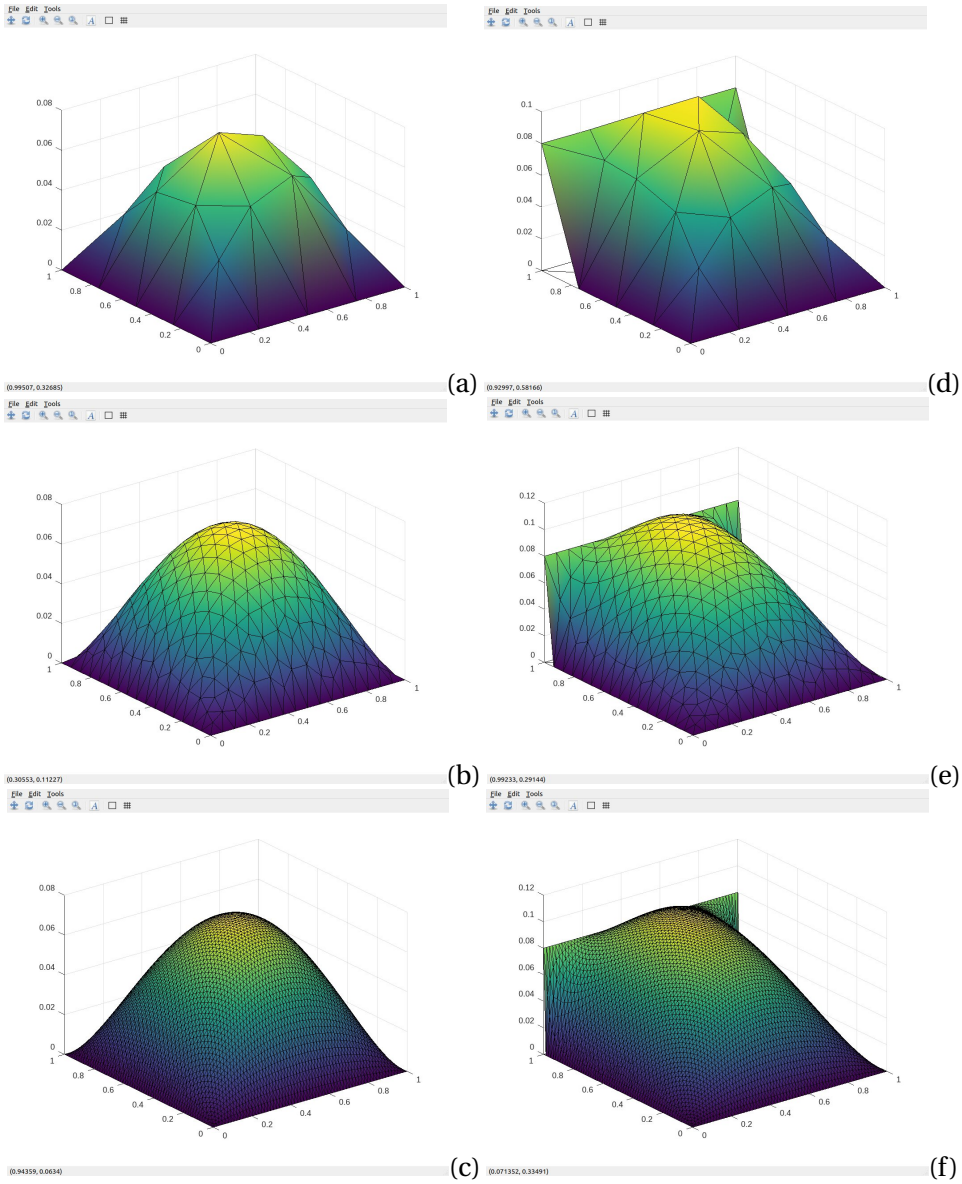


Figure 3.5 Example.

$\delta = 0$	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
M1 ($h = 1/4$)	0.03242	0.001497	0.07488	0.17069	0.18006	0.24431
M2 ($h = 1/8$)	0.03431	0.001641	0.07259	0.17311	0.18524	0.29084
M3 ($h = 1/16$)	0.03491	0.001685	0.07339	0.17412	0.18685	0.31111
M4 ($h = 1/32$)	0.03508	0.001698	0.07363	0.17442	0.18730	0.32606
M5 ($h = 1/64$)	0.03512	0.001701	0.07365	0.17450	0.18742	0.33106
convergence?	YES	YES	YES	YES	YES	YES

In the last row we show the result of a simplified analysis of the sequence of numbers to predict whether it is converging to something or not. It assumes that $Q(h)$ behaves as $q + ch^\beta$ and uses the computed values for the smallest 3 values of h to infer q, c and β. If $\beta > 0$ we say that the sequence is converging.

When $\delta = 0$ the boundary condition is continuous along the boundary. It is observed that all the computed quantities converge, suggesting that Q_1-Q_6 are well defined for the exact solution u.

Of course not all quantities converge with equal speed. It is evident from the table that Q_6 converges much more slowly than all the others.

Let us now turn to the case $\delta = 0.08$. In this case the boundary condition is discontinuous along $\partial\Omega$. **It can be proved that the exact solution to this problem exists and is unique, but it does not belong to** $H^1(\Omega)$, namely, $\|u\|_1 = +\infty$.

From the practical viewpoint, nothing happens. We simply change the imposed value of u_h at some boundary nodes. This does not change the system matrix and thus the linear system is well posed and u_h perfectly defined.

From the theoretical viewpoint, on the other hand, we cannot apply Céa's lemma 3.1 and thus we do not know whether u_h converges or not to u, or in which norm. Most importantly, the quantities of interest may exhibit different behaviors. The table below shows $Q_1 - Q_6$ for several meshes to take a look at what happens.

$\delta = 0.08$	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
M1 ($h = 1/4$)	5.458e-02	6.240e-02	9.962e-02	1.933e-01	2.197e-01	4.224e-01
M2 ($h = 1/8$)	5.494e-02	6.273e-02	1.003e-01	2.050e-01	2.361e-01	7.366e-01
M3 ($h = 1/16$)	5.510e-02	6.288e-02	1.007e-01	2.096e-01	2.490e-01	1.407e+00
M4 ($h = 1/32$)	5.513e-02	6.292e-02	1.008e-01	2.116e-01	2.604e-01	2.764e+00
M5 ($h = 1/64$)	5.514e-02	6.293e-02	1.008e-01	2.125e-01	2.711e-01	5.513e+00
convergence?	YES	YES	YES	YES	**NO**	**NO**

The quantities $Q_1 - Q_4$ are convergent as in the previous case. This is numerical evidence that u belongs to $L^p(\Omega)$ for $p = 1$, 2 and ∞ and that ∇u belongs to $L^1(\Omega)$. As expected since $Q_5(u) = +\infty$, $Q_5(u_h)$ slowly diverges as $h \to 0$. The values obtained for any h are thus meaningless. One should not confuse the small

changes in Q_5 with "mesh convergence" and erroneously infer that "$Q_5(u)$ must be something close to 0.3". The quantity Q_6 diverges severely.

3.6 Summary

- A **finite element method** is defined by a bilinear form $a(\cdot,\cdot)$, a linear form $\ell(\cdot)$, a finite element space $\mathcal{W}_h$ and a treatment of the essential boundary conditions (by interpolation, typically).
- The finite element solution u_h is defined by (3.2). The coercivity condition (3.6), which can easily be checked, guarantees that the linear system is well posed and thus u_h exists and is unique. The condition is not necessary, in fact a weaker version of coercivity is sufficient for existence and uniqueness.
- For u_h to approximate the exact solution u it is **necessary** that the method is **consistent**. This means that the residual $a(u, v_h) - \ell(v_h)$ is identically zero or at least sufficiently small. This condition is usually easy to check and allows to identify the required continuity of functions in $\mathcal{W}_h$ across inter-element boundaries.
- Consider a sequence of meshes $\mathcal{T}_h$ for the problem domain, with $h \to 0$. The corresponding sequence $\{u_h\}$ of numerical solutions then converges to u in the sense of a norm $\|\cdot\|$. For this to happen, **continuity and coercivity** conditions must hold with respect to $\|\cdot\|$.
- All quantities of interest that are continuous with respect to $\|\cdot\|$ then automatically **converge**, and it thus makes sense to consider $Q(u_h)$ as an approximation to $Q(u)$. When h is small enough $|Q(u) - Q(u_h)|$ typically behave as $c\,h^\beta$, β being the order of the approximation.

Chapter 4

Linear Elasticity

In Chapters 1 and 2 we have followed a certain path of which the starting point was always the differential equation and the boundary conditions of a problem. From there, we inferred a variational equation and finally a variational numerical method. The finite element method is the combination of the variational method with some specially crafted subspaces, the finite element spaces.

In this chapter dedicated to linear elasticity we will follow another presentation path, known as a *variational* path. This path does not start from a differential equation, but from a *variational principle*. The variational principle that governs the behavior of many mechanical systems is the **minimization of the energy**. The static deformation of an elastic body, in particular, minimizes the **potential energy**, as we will soon see. The displacement of each tiny piece, be it located near the application of the load or far away from it, is dictated by this principle.

From a variational principle it is possible to deduce a *weak form* of a problem. It is even easier than when we start from the differential equation. Once we arrive at the weak form, the rest of the procedure is as before: Formulate the variational method and propose finite element spaces for it.

4.1 The Variational Problem of Linear Elasticity

The displacement field. Consider the problem sketched in Fig. 4.1. A solid body occupies the domain $\Omega \subset \mathbb{R}^2$, with boundary $\partial\Omega$. Along the **Dirichlet** part of the boundary ($\partial\Omega_D$) a displacement $\mathbf{g}$ is imposed to the particles of the body, while along the **Neumann** part ($\partial\Omega_N$) a distribution of forces $\mathbf{H}$ is applied. In addition, a body force $\mathbf{b}$ loads the body.

Under these conditions, the body will deform. The material particles will change position. This is expressed mathematically by a **displacement field** $\mathbf{u}(\mathbf{x})$, so that the displacement induced by the load on the particle at $\mathbf{x}$ is

$$\mathbf{x} \mapsto \mathbf{x} + \mathbf{u}(\mathbf{x}) \,. \tag{4.1}$$

The underformed domain Ω is called the **reference configuration** of the body, while the deformed domain $\{\mathbf{x}+\mathbf{u}(\mathbf{x}) \mid \mathbf{x} \in \Omega\}$ is called the **deformed configuration**.

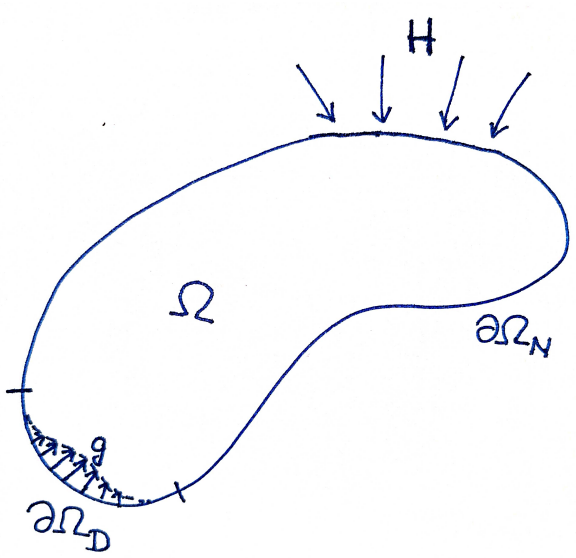


Figure 4.1 A sketch of an elasticity problem. The Dirichlet and Neumann boundaries, $\partial\Omega_D$ and $\partial\Omega_N$, are indicated.

Our quest is to determine the displacement field $\mathbf{u}:\Omega\to\mathbb{R}^2$. The unknown is thus a vector field over Ω, which can also be seen as two unknown functions $u_1(x_1,x_2)$ and $u_2(x_1,x_2)$, since $\mathbf{u}=(u_1,u_2)^T$ or, equivalently,

$$\mathbf{u}(\mathbf{x}) = u_1(\mathbf{x})\,\mathbf{e}_1 + u_2(\mathbf{x})\,\mathbf{e}_2\ .$$

We will seek a physically valid solution $\mathbf{u}$ among displacement fields in a space

$$\mathscr{W} = \{\mathbf{w}:\Omega\to\mathbb{R}^2 \mid \mathbf{w} \text{ is a smooth vector field}\}\ . \tag{4.2}$$

The minimum smoothness required will be discussed further along.

Notice that we already know the value of $\mathbf{u}$ ($=\mathbf{g}=(g_1,g_2)^T$) along $\partial\Omega_D$, but we do not know how this displacement is "distributed" over Ω, or along $\partial\Omega_N$. We define the **trial space** $\mathscr{S}$ as

$$\mathscr{S} = \{\mathbf{w}\in\mathscr{W} \mid \mathbf{w}=\mathbf{g} \text{ on } \partial\Omega_D\}\ . \tag{4.3}$$

The Principle of Minimum Potential Energy. Instead of stating a differential equation to determine $\mathbf{u}$, we will invoke a **variational principle**, in particular, the **principle of minimum potential energy**. The solid is modeled as endowed with an **internal energy** U (also called **strain energy**) which depends solely on the displacement field $\mathbf{u}$. Then the **potential energy** V is defined as

$$V(\mathbf{u}) = U(\mathbf{u}) - \int_\Omega \mathbf{b}\cdot\mathbf{u}\,d\Omega - \int_{\partial\Omega_N}\mathbf{H}\cdot\mathbf{u}\,d\partial\Omega\ . \tag{4.4}$$

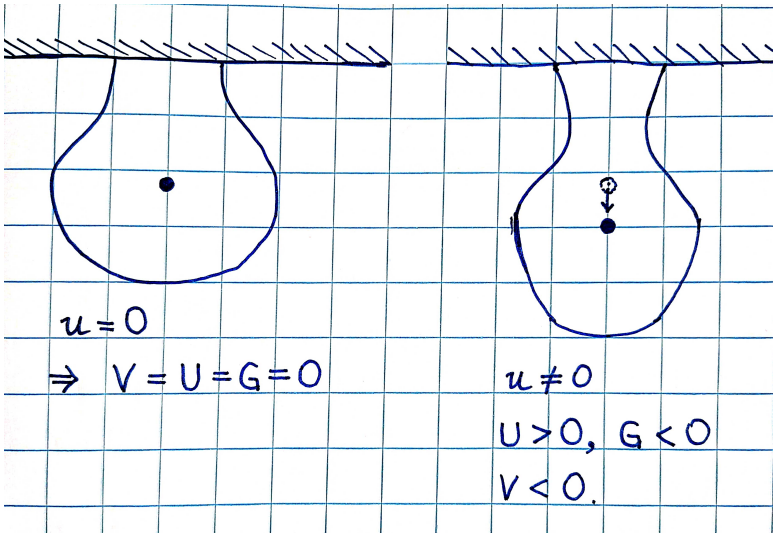


Figure 4.2 An elastic body hanging from the ceiling. On the left we have the reference configuration, which would correspond to gravity being "turned off". On the right we have the equilibrium configuration, which minimizes the potential energy V. Under the load of gravity the body deforms, increasing its strain energy, but this increase is more than compensated by the decrease in gravitational energy, evidenced by the lowering of the center of mass. The equilibrium configuration has less potential energy than the reference one.

The potential energy considers both the strain energy and the work done by the external loads.

The principle reads: *The* **equilibrium solution u** *minimizes the potential energy V among all smooth displacement fields* **w** *that are equal to* **g** *on* $\partial\Omega_D$. In mathematical terms,

$$V(\mathbf{u}) \leq V(\mathbf{w}) , \qquad \text{for all } \mathbf{w} \in \mathscr{S}. \tag{4.5}$$

As an example, consider an elastic body which has a part of its boundary attached to the ceiling (see Fig. 4.2). The body is hanging from its upper fixation. The only applied load is the body's own weight $\mathbf{b} = -\rho\, g\, \mathbf{e}_2$, thus

$$V(\mathbf{u}) = U(\mathbf{u}) + \rho\, g \int_\Omega u_2 \, d\Omega .$$

The actual solution results from a compromise. The second term (which is nothing but the gravitational energy, denoted by G) decreases as the particles of the body displace downwards, but since the body is fixed at the top any downward displacement generates **strains** that make the internal energy increase. The strain energy function U depends on the material of the body, and so does the solution (usually termed **equilibrium solution**).

The small deformation hypothesis (SDH). We adopt here the SDH, which

assumes that the data **g**, **H** and **b** are small enough that **u** and its gradient

$$\nabla \mathbf{u} = \begin{pmatrix} \partial_1 u_1 & \partial_2 u_1 \\ \partial_1 u_2 & \partial_2 u_2 \end{pmatrix} \tag{4.6}$$

are very very small at all points of Ω. So small that the deformed domain essentially coincides with Ω.

Why bother to compute **u** if by hypothesis it is negligible at all points? It happens that solid materials are quite stiff, so that with just tiny displacements they can generate forces that equilibrate significant loads. As an example, consider a 1-m long steel wire with cross-sectional area of 10^{-4} m^2. If it is loaded with 1000 N the maximum displacement is (assuming a Young modulus $E = 2 \times 10^{11}$ Pa)

$$\Delta \ell = \frac{P\ell}{AE} = \frac{1000\,\text{N} \times 1\,\text{m}}{10^{-4}\,\text{m}^2 \times 2 \times 10^{11}\,\text{Pa}} = 5 \times 10^{-5}\,\text{m} = 50\,\text{microns}\,.$$

Consider the decomposition of $\nabla \mathbf{u}$ into symmetric (ε) and anti-symmetric (ω) parts, i.e.,

$$\nabla \mathbf{u} = \varepsilon(\nabla \mathbf{u}) + \omega(\nabla \mathbf{u}) = \frac{1}{2}\left(\nabla \mathbf{u} + \nabla \mathbf{u}^T\right) + \frac{1}{2}\left(\nabla \mathbf{u} - \nabla \mathbf{u}^T\right)\,. \tag{4.7}$$

Under the SDH, it can be shown that the tensor (or matrix) ε measures the **local deformation** (or **strain**) of the solid, while ω measures the **local rotation**.

The strain energy of a linearly elastic body under the SDH. If the material is **isotropic**, the strain energy takes the form

$$U(\mathbf{u}) = \int_\Omega \frac{E}{2(1+\nu)}\left(\varepsilon(\nabla \mathbf{u}) : \varepsilon(\nabla \mathbf{u}) + \frac{\nu}{1-2\nu}(\text{div}\,\mathbf{u})^2\right)\,d\Omega\,, \tag{4.8}$$

where E is Young's modulus, ν is Poisson's ratio, $\text{div}\,\mathbf{u} = \partial_1 u_1 + \partial_2 u_2 = \varepsilon_{11} + \varepsilon_{22}$, and ":" stands to the double contraction of two tensors, or equivalently the Frobenius product of two matrices:

$$\mathbf{A} : \mathbf{B} = \sum_{i,j} A_{ij} B_{ij}\,.$$

The value of Young's modulus is always a positive number, and Poisson's ratio spans the range $-1 \le \nu \le 0.5$. For example, metals typically have Poisson's ratios between 0.1 and 0.4. The integrand in (4.8) is the **strain energy density** at each point **x**, which is non-negative, and it is zero if and only if $\nabla \mathbf{u}(\mathbf{x})$ is antisymmetric. Therefore, the strain energy is always non-negative, and it is minimal if $\nabla \mathbf{u}(\mathbf{x})$ is anti-symmetric for all $\mathbf{x} \in \Omega$, or a small rigid body rotation.

If the material is not isotropic the corresponding expression is less appealing. The constitutive coefficients E and ν turn into a fourth-order array of coefficients C_{ijkl}, and then

$$U(\mathbf{u}) = \frac{1}{2}\int_\Omega \sum_{i,j,k,l} C_{ijkl}\,\varepsilon_{ij}(\nabla \mathbf{u})\,\varepsilon_{kl}(\nabla \mathbf{u})\,d\Omega\,. \tag{4.9}$$

In the following we will restrict the discussion to isotropic materials, so that the two coefficients E and ν (which may depend on **x**) characterize the elastic properties of the material.

Problem 4.1 ("Primal" variational form of the isotropic linear elasticity problem)**.** *Given the domain* Ω *and the data* E, ν, $\mathbf{b}$, $\mathbf{g}$ *and* $\mathbf{H}$, *determine the smooth vector field* $\mathbf{u} \in \mathscr{S}$ *that minimizes* V *over* $\mathscr{S}$, *where for a generic* $\mathbf{w} \in \mathscr{W}$ *the potential energy function is*

$$V(\mathbf{w}) = \int_\Omega \frac{E}{2(1+\nu)} \left(\varepsilon(\nabla \mathbf{w}) : \varepsilon(\nabla \mathbf{w}) + \frac{\nu}{1-2\nu} (div\ \mathbf{w})^2 \right) d\Omega \\ - \int_\Omega \mathbf{b} \cdot \mathbf{w}\, d\Omega - \int_{\partial\Omega_N} \mathbf{H} \cdot \mathbf{w}\, d\partial\Omega\,. \quad (4.10)$$

By replacing the first term above by (4.9) one obtains the primal variational form for anisotropic materials.

Example 4.1 A sphere under pressure. What is the deformation of a spherical homogeneous elastic body when a radial force $\mathbf{H} = -p\,\check{\mathbf{e}}_r$ is applied to its surface? The displacement will be radially symmetric, i.e.,

$$\mathbf{u}(\mathbf{x}) = \varphi(r)\,\check{\mathbf{e}}_r.$$

From the expression of the gradient in spherical coordinates we know that

$$\varepsilon(\nabla \mathbf{u}) : \varepsilon(\nabla \mathbf{u}) = \varphi'(r)^2 + 2\frac{\varphi(r)^2}{r^2}$$

and that

$$\operatorname{div} \mathbf{u} = \varphi'(r) + 2\,\frac{\varphi(r)}{r}\,.$$

The physical restriction $\mathbf{u}(\mathbf{0}) = \mathbf{0}$ translates into $\varphi(0) = 0$. The equilibrium displacement of the sphere will thus be given by the function φ that minimizes, over the set of smooth functions

$$\mathscr{S} = \mathscr{V} = \left\{ \varphi : [0, R] \to \mathbb{R} \mid \varphi(0) = 0 \right\},$$

the potential energy

$$V(\varphi) \;=\; \frac{E}{2(1+\nu)} \int_0^R \left(\varphi'(r)^2 + 2\frac{\varphi(r)^2}{r^2} + \frac{\nu}{1-2\nu} \left(\varphi'(r) + 2\,\frac{\varphi(r)}{r} \right)^2 \right) 4\pi r^2 dr \;+ \\ +4\pi p R^2 \varphi(R)\,. \quad (4.11)$$

It turns out that the exact minimizer is of the form $\varphi(r) = A\,r$ for some $A \in \mathbb{R}$ that depends on p. If we assume that $\mathscr{S}$ only contains such functions, then V becomes a function of A, that is

$$V = \gamma A^2 + 4\pi p R^3\, A, \qquad \text{with} \qquad \gamma = \frac{2\pi E R^3}{1+\nu} \left(1 + \frac{3\nu}{1-2\nu} \right) = \frac{2\pi E R^3}{1-2\nu}\,.$$

The minimum takes place for

$$A = -\frac{4\pi p R^3}{2\gamma} = -\frac{1-2\nu}{E}\, p\,. \tag{4.12}$$

We have thus our first solution of a linear elastic problem. The displacement field is

$$\mathbf{u}(\mathbf{x}) = A\, r\, \check{\mathbf{e}}_r = A\mathbf{x} = -\frac{1-2\nu}{E}\, p\, \mathbf{x}.$$

Notice that if $\nu = \frac{1}{2}$ then $A = 0$ and thus $\mathbf{u} = \mathbf{0}$ at all points. The sphere does not contract under applied pressure. The limit $\nu \to \frac{1}{2}$ is called the "incompressible limit."

We cannot yet prove that this is the exact solution, but we can at least confirm that the polynomial $A\,r$ minimizes V over all **quadratic** polynomials. For this, we take $\varphi(r) = A\,r + B\,r^2$, with unknown coefficients A and B. Inserting this φ into (4.11) we obtain V as a function of A and B, namely

$$V(A,B) = \frac{2\pi E}{1-2\nu}\left[R^3\,A^2 + 2R^4\,A\,B + \frac{6+4\nu}{5+5\nu}\,R^5\,B^2\right] + 4\pi R^3 p\,(A + R\,B)$$

By equating $\partial V/\partial A = \partial V/\partial B = 0$ you can check that the minimum takes place when A takes the value given in (4.12) and $B = 0$.

Remark: If we consider a 2D sphere (a circle), then

$$\varepsilon(\nabla\mathbf{u}) : \varepsilon(\nabla\mathbf{u}) = \varphi'(r)^2 + \frac{\varphi(r)^2}{r^2} \quad \text{and} \quad \operatorname{div}\mathbf{u} = \varphi'(r) + \frac{\varphi(r)}{r}\,.$$

This modifies the expression of $V(A)$ to

$$V = \frac{\pi E R^2}{(1+\nu)(1-2\nu)}A^2 + 2\pi p R^2\, A,$$

and the value of A that corresponds to the minimum is

$$A = -\frac{(1+\nu)(1-2\nu)}{E}\, p\,. \tag{4.13}$$

Under the same pressure, a circle deforms less than a sphere.

4.2 From the Variational Form to the Weak Form

How to obtain a weak form when starting from a variational principle? It is quite straightforward. The procedure is based on the following rather abstract theorem.

Theorem 4.1. *Let $\mathcal{W}$ be a* **vector space** *(it could be of functions, of vector fields, etc.), and let $\mathcal{S}$ be an affine subspace of $\mathcal{W}$. The* **direction** *of $\mathcal{S}$ is denoted by $\mathcal{V}$. Assume that:*

a) *An energy function V is defined on $\mathscr{W}$ which can be written as*

$$V(w) = \frac{1}{2}a(w,w) - \ell(w) \tag{4.14}$$

where a is a symmetric bilinear form satisfying that

$$a(v,v) > 0 \qquad \forall v \in \mathscr{V}, \quad v \neq 0,$$

and ℓ is a linear form.

b) *There exists a minimizer u of V over $\mathscr{S}$. Precisely, there exists $u \in \mathscr{S}$ satisfying*

$$V(u) \leq V(w), \qquad \forall w \in \mathscr{S}\,. \tag{4.15}$$

Then, u is the unique minimizer in $\mathscr{S}$ (i.e., $V(u) < V(w)$, $\forall w \neq u$). In addition, u is also the unique element of $\mathscr{S}$ satisfying

$$a(u,v) = \ell(v) \qquad \forall v \in \mathscr{V}\,. \tag{4.16}$$

Proof. First let us prove that u necessarily satisfies (4.16). For this, assume that there exists some particular $0 \neq v \in \mathscr{V}$ for which

$$a(u,v) - \ell(v) = \beta \neq 0\,.$$

We will show that then u is not a minimizer of V over $\mathscr{S}$. Let us define $\alpha = a(v,v) > 0$ (because of hypothesis (a)) and

$$w = u - \frac{\beta}{\alpha}v\,.$$

Using the linearity of $a(\cdot,\cdot)$ and ℓ and the symmetry of $a(\cdot,\cdot)$ we see that

$$\begin{aligned} V(w) &= \frac{1}{2}a\left(u - \frac{\beta}{\alpha}v, u - \frac{\beta}{\alpha}v\right) - \ell\left(u - \frac{\beta}{\alpha}v\right) \\ &= \frac{1}{2}a(u,u) - \ell(u) - \frac{\beta}{\alpha}\underbrace{(a(u,v) - \ell(v))}_{\beta} + \frac{\beta^2}{2\alpha^2}\underbrace{a(v,v)}_{\alpha} \\ &= V(u) - \frac{\beta^2}{2\alpha} \\ &< V(u)\,. \end{aligned}$$

This proves (4.16). Now assume that there exists another element of $\mathscr{S}$, let us call it $\overline{u}$, that also satisfies $a(\overline{u},v) = \ell(v)$ for all $v \in \mathscr{V}$. Then,

$$\begin{aligned} a(u-\overline{u}, u-\overline{u}) &= a(u, \underbrace{u-\overline{u}}_{\in\mathscr{V}}) - a(\overline{u}, \underbrace{u-\overline{u}}_{\in\mathscr{V}}) \\ &= \ell(u-\overline{u}) - \ell(u-\overline{u}) \qquad = 0\,. \end{aligned}$$

According to hypothesis (a) this implies $u - \overline{u} = 0$ and thus $u = \overline{u}$.

This last argument also proves that u is the unique minimizer, since the existence of two minimizers would imply the existence of two solutions of (4.16) and that has been shown to be impossible. □

This theorem, though elementary, has interesting consequences. Notice that very little is said about the space $\mathcal{W}$. It could be finite or infinite dimensional, for example. The bilinear and linear forms are not required to be continuous. The hypotheses on the spaces involved are very weak, but on the other hand we make the strong assumption that a unique minimizer exists. Sometimes physical reasons make us believe that a unique minimum exists, though a mathematical proof could be unavailable. Under this assumption, the theorem provides us with **a variational equation for the problem**, namely $a(u,v) = \ell(v), \quad \forall v \in \mathcal{V}$.

Applying this theorem to Problem (4.1) we obtain:

Problem 4.2 (Weak form of the Isotropic Linear Elasticity Problem)**.** *Given the domain* Ω *and the data* E, ν, $\mathbf{b}$, $\mathbf{g}$ *and* $\mathbf{H}$*, determine the smooth vector field* $\mathbf{u} \in \mathcal{S}$ *(i.e.,* $\mathbf{u} = \mathbf{g}$ *on* $\partial\Omega_D$*) such that*

$$a(\mathbf{u},\mathbf{v}) = \ell(\mathbf{v}) \tag{4.17}$$

for all $\mathbf{v} \in \mathcal{V}$*, where*

$$a(\mathbf{w},\mathbf{v}) = \int_\Omega \frac{E}{1+\nu}\left(\varepsilon(\nabla\mathbf{w}) : \varepsilon(\nabla\mathbf{v}) + \frac{\nu}{1-2\nu} div\,\mathbf{w}\; div\,\mathbf{v}\right) d\Omega, \tag{4.18}$$

$$\ell(\mathbf{v}) = \int_\Omega \mathbf{b}\cdot\mathbf{v}\, d\Omega + \int_{\partial\Omega_N} \mathbf{H}\cdot\mathbf{v}\, d\partial\Omega\,. \tag{4.19}$$

and

$$\mathcal{V} = \{\mathbf{v} \in \mathcal{W} \mid \mathbf{v} = \mathbf{0} \text{ on } \partial\Omega_D\}\,. \tag{4.20}$$

We should check that the hypotheses of Theorem 4.1 indeed hold true. It is readily seen that $V(\mathbf{w}) = \frac{1}{2}a(\mathbf{w},\mathbf{w}) - \ell(\mathbf{w})$. A little more subtle is to prove that $a(\mathbf{v},\mathbf{v}) > 0$ for all $\mathbf{v} \in \mathcal{V}$, $\mathbf{v} \neq \mathbf{0}$. We do this below when we study its coercivity, and it is true for $E > 0$ and $-1 \leq \nu \leq 0.5$, the range of values of these elastic moduli. It also requires that $\mathcal{V}$ does not contain any *rigid mode* (infinitesimal translations/rotations, which have $\varepsilon = 0$); to be seen during the study of coercivity as well.

Then, the theorem tells us that Problem 4.2 has **as unique solution** the displacement field $\mathbf{u}$ that minimizes the potential energy V (assumed to exist).

The stress field. The integrand in (4.18) can also be written as

$$\sigma(\nabla\mathbf{u}) : \varepsilon(\nabla\mathbf{v})$$

where σ is the **Cauchy stress tensor**, or simply **stress tensor**,

$$\sigma = \frac{E}{1+\nu}\,\varepsilon(\nabla\mathbf{u}) + \frac{E\nu}{(1+\nu)(1-2\nu)}\,(\underbrace{\operatorname{div}\mathbf{u}}_{=\mathbf{I}:\nabla\mathbf{u}})\,\mathbf{I}\,. \tag{4.21}$$

In many cases the elasticity problem is solved mainly looking for the stress field over the body, since too high stresses may lead to the failure of the material.

An alternative set of elastic constants are the Lamé constants

$$\lambda = \frac{E}{(1+\nu)(1-2\nu)} \qquad \text{and} \qquad \mu = \frac{E}{2(1+\nu)},$$

in terms of which the stress tensor is obtained as

$$\sigma = \lambda \mathrm{tr}(\varepsilon(\mathbf{u}))\mathbf{I} + 2\mu\varepsilon(\nabla\mathbf{u}),$$

where

$$\mathrm{tr}(\varepsilon(\mathbf{u})) = \varepsilon(\mathbf{u})_{11} + \varepsilon(\mathbf{u})_{22} = \mathrm{div}\,\mathbf{u}$$

is the **trace** of the strain $\varepsilon(\mathbf{u})$.

Example 4.2 The stress field inside a sphere under pressure. We saw in Example 4.1 that the equilibrium displacement field of a 2D sphere (circle) under uniform pressure is

$$\mathbf{u}(\mathbf{x}) = A\,\mathbf{x}, \quad \text{where} \quad A = -\frac{(1+\nu)(1-2\nu)p}{E}$$

and $\mathbf{x}$ has the center of the circle as origin. This equation is intrinsic, valid for all coordinate systems. We can thus express it in **Cartesian** coordinates $x_1 - x_2$ so that $\mathbf{x} = (x_1, x_2)^T$. Then $u_1 = Ax_1$ and $u_2 = Ax_2$ and thus

$$\nabla\mathbf{u} = \varepsilon(\nabla\mathbf{u}) = \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix} = A\mathbf{I}, \quad \omega(\nabla\mathbf{u}) = 0, \quad \mathrm{div}\ \mathbf{u} = 2A.$$

Inserting these values into (4.21) we obtain the corresponding stress field:

$$\sigma = \frac{EA}{1+\nu}\mathbf{I} + \frac{2E\nu A}{(1+\nu)(1-2\nu)}\mathbf{I} = \frac{EA}{(1+\nu)(1-2\nu)}\mathbf{I} = -p\,\mathbf{I} = \begin{pmatrix} -p & 0 \\ 0 & -p \end{pmatrix}$$

The stress field is **homogeneous**. It does not depend on $\mathbf{x}$. It is called **spherical** (or **hydrostatic**) because at all points it is a multiple of the identity tensor/matrix.

The Euler-Lagrange equations. The weak form is all we need to set up a variational method to approximate the exact solution $\mathbf{u}$. Let us however briefly talk about the Euler-Lagrange equations of the problem, and formulate the strong form. As usual, it is obtained by integrating by parts the weak form. Since σ is symmetric, it holds that $\sigma : \varepsilon(\nabla\mathbf{v}) = \sigma : \nabla\mathbf{v}$

$$\begin{aligned}\int_\Omega \sigma : \varepsilon(\nabla\mathbf{v})\, d\Omega = \int_\Omega \sigma : \nabla\mathbf{v}\, d\Omega &= \sum_{i,j}\int_\Omega \sigma_{ij}\frac{\partial u_i}{\partial x_j}\, d\Omega \\ &\sum_{i,j}\left[\int_{\partial\Omega}\sigma_{ij}\check{n}_j u_i\, \partial\Omega - \int_\Omega \frac{\partial\sigma_{ij}}{\partial x_j}v_i\, d\Omega\right] \\ &= \int_{\partial\Omega}(\sigma\cdot\check{\mathbf{n}})\cdot\mathbf{v}\, d\partial\Omega - \int_\Omega (\mathrm{div}\ \sigma)\cdot\mathbf{v}\, d\Omega\,.\end{aligned}$$

From this, remembering that $\mathbf{v} = 0$ on $\partial\Omega_D$, we arrive at

$$0 = a(\mathbf{u},\mathbf{v}) - \ell(\mathbf{v}) = \int_\Omega (-\mathrm{div}\ \sigma - \mathbf{b})\cdot\mathbf{v}\, d\Omega + \int_{\partial\Omega_N}(\mathbf{H} - \sigma\cdot\check{\mathbf{n}})\cdot\mathbf{v}\, d\partial\Omega\,,$$

for all $\mathbf{v} \in \mathcal{V}$. From this we conclude that $\mathbf{u}$ is also the solution of the following differential problem.

Problem 4.3 (Strong form of the Isotropic Linear Elasticity Problem)**.** *Given the same data as in Problems 4.1 and 4.2, find a smooth vector field* **u** *satisfying*

$$\begin{aligned} div\ \sigma(\nabla\mathbf{u}) + \mathbf{b} &= 0 \quad in\ \Omega, && (4.22)\\ \mathbf{u} &= \mathbf{g} \quad on\ \partial\Omega_D, && (4.23)\\ \sigma(\nabla\mathbf{u})\cdot\check{\mathbf{n}} &= \mathbf{H} \quad on\ \partial\Omega_N, && (4.24) \end{aligned}$$

where σ *is defined by (4.21).*

The divergence of the stress field is a vector field with Cartesian components

$$(\mathrm{div}\sigma)_i = \sum_{j=1}^{2} \frac{\partial \sigma_{ij}}{\partial x_j} = \frac{\partial \sigma_{i1}}{\partial x_1} + \frac{\partial \sigma_{i2}}{\partial x_2}$$

for $i = 1, 2$.

Equation (4.22) is also known as **equation of static equilibrium**. It expresses the local equilibrium of forces at each point of the domain, irrespective of the material being linearly elastic or not. Of course, if the material is not linearly elastic the expression for σ is different from (4.21).

It is important to internalize that Problems 4.1, 4.2 and 4.3 are essentially equivalent. Each one of them totally determines **u**.

Examples:

4.3 **The exact solution of the problem of a sphere under pressure.** It is easy to verify that the stress field $\sigma = -p\mathbf{I}$ computed in Example 4.2 is a solution to Problem 4.3. In fact, since σ is independent of **x**, div $\sigma = 0$, which is consistent with (4.22) because $\mathbf{b} = 0$. Also, since the boundary condition is $\mathbf{H} = -p\check{\mathbf{e}}_r$ and $\check{\mathbf{n}} = \check{\mathbf{e}}_r$, it follows that $\sigma\cdot\check{\mathbf{n}} = \mathbf{H}$ all over the boundary. Then the displacement field **u** calculated in Example 4.1 is indeed the unique solution (up to a rigid motion) of the problem of an isotropic linear elastic sphere under pressure.

4.4 **The exact solution of the problem of a rectangle under pressure.** Assume that the body subject to a uniform pressure p over its surface is the rectangle $\Omega = [-\frac{W}{2}, \frac{W}{2}] \times [-\frac{H}{2}, \frac{H}{2}]$. The force imposed by the pressure is $\mathbf{H} = -p\check{\mathbf{n}}$ at all points, so that **H** takes the value $-p\check{\mathbf{e}}_1$ at the east boundary, $p\check{\mathbf{e}}_1$ at the west one, $-p\check{\mathbf{e}}_2$ at the north one, and $p\check{\mathbf{e}}_2$ at the south one. The same displacement and stress fields (though now defined in the rectangular domain Ω) that solve the elastic problem for the sphere also solve it for the rectangle, any rectangle. In fact, **for any shape**.

The space of functions for elasticity. The space of candidate displacement fields for elasticity $\mathcal{W}$ should satisfy some minimal conditions. First, the potential energy $V(\mathbf{u})$ of any displacement field $u \in \mathcal{W}$ should be finite. Second, unless we consider situations in which fractures may appear (and additional modeling is needed), the displacement fields should not be discontinuous across curves with positive length in the domain of the problem (in 3D it should be across surfaces with positive area); pointwise discontinuities are still acceptable.

A space in which all of these requirements are met is

$$\mathcal{W} = \mathbf{H}^1(\Omega) = \left\{\mathbf{w} = (w_1, w_2)^T : \Omega \to \mathbb{R}^2 \mid w_1 \in H^1(\Omega),\ w_2 \in H^1(\Omega)\right\}, \tag{4.25}$$

where $H^1(\Omega)$ is the space of functions over Ω with a finite H^1-norm[1] (see A.11), namely, if $w \in H^1(\Omega)$ then

$$\|w\|_{1,2} = \left[\int_\Omega w^2 + |\nabla w|^2 \, d\Omega\right]^{1/2} < +\infty. \tag{4.26}$$

So, $\mathcal{W}$ is a space of vector fields in which each component is a function in $H^1(\Omega)$. A norm on $\mathcal{W}$ is

$$\|\mathbf{w}\|_{1,2} = \sqrt{\|w_1\|_{1,2}^2 + \|w_2\|_{1,2}^2}\,. \tag{4.27}$$

In fact, $a(\cdot,\cdot)$ has the necessary properties of continuity in $\mathbf{H}^1(\Omega)$, and coercivity on $\mathcal{V}$, for $-1 < \nu < 0.5$.

Continuity of the Bilinear Form. For $-1 < \nu < 0.5$, the bilinear form satisfies that

$$|a(\mathbf{w},\mathbf{v})| \le M, \|\mathbf{w}\|_{1,2}\ \|\mathbf{v}\|_{1,2} \tag{4.28}$$

for any two elements $\mathbf{w}$ and $\mathbf{v}$ of $\mathcal{W}$, where

$$M = \sup_\Omega \max\left\{\frac{E}{1+\nu}, \frac{E}{1-2\nu}\right\}. \tag{4.29}$$

The constant M tends to infinity when E tends to infinity, or ν tends to $\frac{1}{2}$, somewhere over the domain.

This is a general result on continuity that seems to differ from the requirement in Céa's Lemma, (3.4), but it implies it if we choose $\mathcal{W}_h \subset \mathcal{W}$. Since then $\mathcal{S}_h, \mathcal{V}_h \subseteq \mathcal{W}_h$, and (4.28) holds for any $\mathbf{w}$ and $\mathbf{v}$ of $\mathcal{W}$, we conclude that

$$|a(\mathbf{u}-\mathbf{w}_h, \mathbf{v}_h)| \le M, \|\mathbf{u}_h - \mathbf{w}_h\|_{1,2}\ \|\mathbf{v}_h\|_{1,2} \qquad \forall \mathbf{v}_h \in \mathcal{V}_h, \forall \mathbf{w}_h \in \mathcal{S}_h. \tag{4.30}$$

Continuity of the bilinear form, (4.28)

[1] This is almost the definition of H^1; we also need to require that the function should have *weak derivatives*.

To see the continuity, you can verify that for any $\mathbf{v},\mathbf{w}\in\mathscr{W}$,

$$\frac{E}{(1+\nu)}\left(\varepsilon(\nabla\mathbf{w}):\varepsilon(\nabla\mathbf{v})+\frac{\nu}{1-2\nu}\operatorname{div}\mathbf{w}\operatorname{div}\mathbf{v}\right)=\bar{\varepsilon}(\nabla\mathbf{w})^{\mathsf{T}}\cdot D\cdot\bar{\varepsilon}(\nabla\mathbf{v}) \tag{4.31}$$

where,

$$\bar{\varepsilon}(\nabla\mathbf{v})=\begin{bmatrix}\varepsilon_{11}(\nabla\mathbf{v})\\ \varepsilon_{22}(\nabla\mathbf{v})\\ \varepsilon_{33}(\nabla\mathbf{v})\\ \sqrt{2}\varepsilon_{12}(\nabla\mathbf{v})\\ \sqrt{2}\varepsilon_{13}(\nabla\mathbf{v})\\ \sqrt{2}\varepsilon_{23}(\nabla\mathbf{v})\end{bmatrix},\qquad D=\frac{E}{1+\nu}\begin{bmatrix}1+\frac{\nu}{1-2\nu} & \frac{\nu}{1-2\nu} & \frac{\nu}{1-2\nu} & 0 & 0 & 0\\ \frac{\nu}{1-2\nu} & 1+\frac{\nu}{1-2\nu} & \frac{\nu}{1-2\nu} & 0 & 0 & 0\\ \frac{\nu}{1-2\nu} & \frac{\nu}{1-2\nu} & 1+\frac{\nu}{1-2\nu} & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\end{bmatrix}.$$

The matrix D is positive definite, since its eigenvalues are

$$\frac{E}{1+\nu}\left\{\frac{1+\nu}{1-2\nu},1,1,1,1,1\right\}. \tag{4.32}$$

Therefore, denoting the eigenvalues by $\lambda_1,\ldots,\lambda_6$, and by $\|\varepsilon\|=\sqrt{\varepsilon:\varepsilon}$ the Frobenius norm, we have

$$\begin{aligned}
|\bar{\varepsilon}(\nabla\mathbf{v})^{\mathsf{T}}\cdot D\cdot\bar{\varepsilon}(\nabla\mathbf{w})| &= \left|\sum_{i=1}^{6}\bar{\varepsilon}_i(\nabla\mathbf{v})\bar{\varepsilon}_i(\nabla\mathbf{w})\lambda_i\right| && (4.33)\\
&\le \sum_{i=1}^{6}|\bar{\varepsilon}_i(\nabla\mathbf{v})||\bar{\varepsilon}_i(\nabla\mathbf{w})||\lambda_i|\\
&\le \max_i|\lambda_i|\sum_{i=1}^{6}|\bar{\varepsilon}(\nabla\mathbf{v})||\bar{\varepsilon}(\nabla\mathbf{w})|\\
&\le \max_i|\lambda_i|\|\bar{\varepsilon}(\nabla\mathbf{v})\|\|\bar{\varepsilon}(\nabla\mathbf{w})\| && \text{Cauchy-Schwartz.}\\
&= \max\left\{\frac{E}{1+\nu},\frac{E}{1-2\nu}\right\}\|\varepsilon(\nabla\mathbf{v})\|\|\varepsilon(\nabla\mathbf{w})\| && \text{since } \|\bar{\varepsilon}\|=\|\varepsilon\|\\
&\le \max\left\{\frac{E}{1+\nu},\frac{E}{1-2\nu}\right\}\|\nabla\mathbf{v}\|\|\nabla\mathbf{w}\| && \text{Triangle inequality.}
\end{aligned}$$

In the last line we used that

$$\|\varepsilon(\nabla\mathbf{w})\|=\frac{1}{2}\|\nabla\mathbf{w}+\nabla\mathbf{w}^T\|\le\|\nabla\mathbf{w}\|$$

from the triangle inequality. The version of Cauchy-Schwartz inequality we used here is one for vectors in $\mathbb{R}^n$, instead of functions: if $a,b\in\mathbb{R}^n$, then

$$|a\cdot b|\le\|a\|\|b\|. \tag{4.34}$$

Then,

$$\begin{aligned}
|a(\mathbf{w},\mathbf{v})| &= \int_\Omega\frac{E}{1+\nu}\left(\varepsilon(\nabla\mathbf{w}):\varepsilon(\nabla\mathbf{v})+\frac{\nu}{1-2\nu}\operatorname{div}\mathbf{w}\operatorname{div}\mathbf{v}\right)d\Omega,\\
&\le \int_\Omega\max\left\{\frac{E}{1+\nu},\frac{E}{1-2\nu}\right\}\|\nabla\mathbf{v}\|\|\nabla\mathbf{w}\|d\Omega\\
&\le \sup_\Omega\max\left\{\frac{E}{1+\nu},\frac{E}{1-2\nu}\right\}\|\nabla\mathbf{w}\|_{0,\Omega}\,\|\nabla\mathbf{v}\|_{0,\Omega},\\
&\le M\|\mathbf{w}\|_{1,2}\,\|\mathbf{v}\|_{1,2}.
\end{aligned}$$

Coercivity. Well-posed elasticity problems have Dirichlet boundary conditions that preclude rigid motions. In fact, **rigid motions are automatically excluded from $\mathcal{V}$ if the length (measure) of $\partial\Omega_D$ is strictly positive.** Under this assumption, it is not difficult to prove that $a(\cdot,\cdot)$ is coercive on $\mathcal{V}$, i.e.,

Lemma 4.1. *There exists $c_{\mathcal{V}} > 0$ such that*

$$a(\mathbf{v},\mathbf{v}) \geq c_{\mathcal{V}} \|\mathbf{v}\|_{1,2}^2 \qquad \forall\, \mathbf{v} \in \mathcal{V}. \tag{4.35}$$

Coercivity of the bilinear form, Lemma 4.1

To prove coercivity, we need to appeal to a classical result known as **Korn's inequality**. The version that we use here is:

Korn's inequality: There exists $C_K > 0$ such that, for all $\mathbf{v} \in \mathbf{H}^1(\Omega)$ satisfying $\mathbf{v} = 0$ on $\partial\Omega_D$ (which has to have positive length/measure),

$$\int_\Omega \|\varepsilon(\nabla\mathbf{v})\|^2 \, d\Omega \;\geq\; C_K \|\mathbf{v}\|_{1,2}^2. \tag{4.36}$$

We can now proceed with the proof of Lemma 4.1. Using (4.31) and the eigenvalues of the matrix D, (4.32), we can write

$$\begin{aligned}
\bar{\varepsilon}(\nabla\mathbf{v})^{\mathsf{T}} \cdot D \cdot \bar{\varepsilon}(\nabla\mathbf{v}) &\geq \min\left\{\frac{E}{1+\nu}, \frac{E}{1-2\nu}\right\} \bar{\varepsilon}(\nabla\mathbf{v})^{\mathsf{T}} \cdot \bar{\varepsilon}(\nabla\mathbf{v})^{\mathsf{T}} \\
&\geq \frac{E}{3} \bar{\varepsilon}(\nabla\mathbf{v})^{\mathsf{T}} \cdot \bar{\varepsilon}(\nabla\mathbf{v})^{\mathsf{T}} \\
&= \frac{E}{3} \varepsilon(\nabla\mathbf{v}) : \varepsilon(\nabla\mathbf{v}),
\end{aligned}$$

Remembering that $\|\varepsilon\| = \sqrt{\varepsilon : \varepsilon}$, we have that

$$\begin{aligned}
a(\mathbf{v},\mathbf{v}) &= \int_\Omega \frac{E}{(1+\nu)} \left(\varepsilon(\nabla\mathbf{v}) : \varepsilon(\nabla\mathbf{v}) + \frac{\nu}{1-2\nu} (\operatorname{div} \mathbf{v})^2\right) d\Omega \\
&\geq \frac{E}{3} \int_\Omega \|\varepsilon(\nabla\mathbf{v})\|^2 \, d\Omega \\
&\geq \frac{E}{3} C_K \|\mathbf{v}\|_{1,2}^2 \qquad \text{Korn's inequality, (4.36).}
\end{aligned}$$

4.3 Variational Numerical Method

The variational numerical method to approximate $\mathbf{u}$ is built directly from Problem 4.2, as we have done in all previous chapters. Selecting a finite dimensional space $\mathcal{W}_h$ and constructing $\mathcal{S}_h$ and $\mathcal{V}_h$ as earlier, it reads:

Variational numerical method: *Find $\mathbf{u}_h \in \mathcal{S}_h$ such that*

$$a(\mathbf{u}_h,\mathbf{v}_h) = \ell(\mathbf{v}_h) \tag{4.37}$$

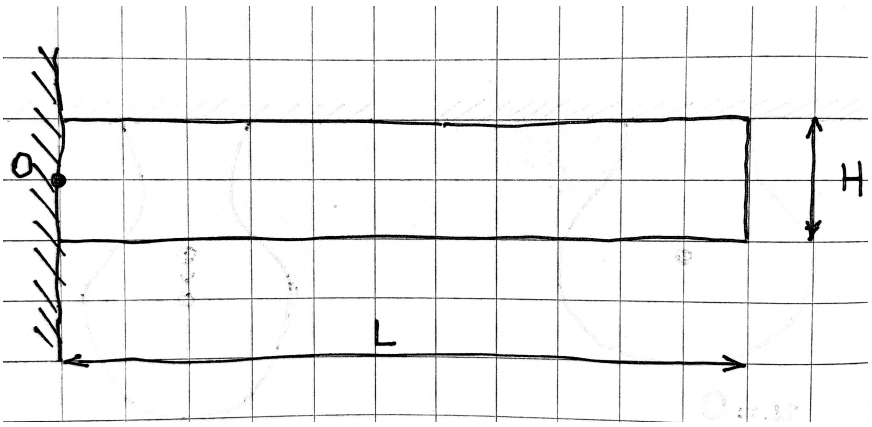


Figure 4.3 Sketch of a 2D cantilever elastic bar under its own weight.

for all $\mathbf{v}_h \in \mathcal{V}_h$*, where* $a(\cdot,\cdot)$ *and* $\ell(\cdot)$ *are given by (4.17) and (4.19), respectively.*

Expressed in this abstract form the only difference with our previous encounters with the variational methods is the bold characters used to denote the solution $\mathbf{u}_h$ and the test function $\mathbf{v}_h$. This is just a purely notational innovation we have introduced to remind us that $\mathcal{S}_h$ and $\mathcal{V}_h$ are spaces of **vector fields**.

For consistency, again, functions in $\mathcal{V}_h$ should be continuous vector fields, and hence we set $\mathcal{W}_h$ to be a space of continuous vector fields.

Before describing suitable finite element spaces and bases for the variational method, let us illustrate how it works when using a basis of **global polynomials**, which certainly are continuous.

Example 4.5 (A cantilever rectangular bar under its own weight) Consider the 2D rectangular bar of Figure 4.3 ($\Omega = [0, L] \times [-\frac{H}{2}, \frac{H}{2}]$), which is fixed to the rigid wall along $x_1 = 0$ and subject to its own body weight $\mathbf{b} = (0, -\rho g)^\mathsf{T}$. There are no other loads.

As $\mathcal{W}_h$ we take polynomials (in the two variables x_1 and x_2) of degree up to k_1 for u_1 and up to k_2 for u_2. In other words,

$$\mathcal{W}_h = \left\{ \mathbf{w}_h : \Omega \to \mathbb{R}^2 \mid w_{h1} \in \mathbb{P}_{k_1},\ w_{h2} \in \mathbb{P}_{k_2} \right\}$$

Any $\mathbf{w}_h \in \mathcal{W}_h$, taking $k_1 = 1$, $k_2 = 2$ as an example, can be written as

$$\mathbf{w}_h = \begin{pmatrix} c_1 + c_2 x_1 + c_3 x_2 \\ \\ c_4 + c_5 x_1 + c_6 x_2 + c_7 x_1^2 + c_8 x_1 x_2 + c_9 x_2^2 \end{pmatrix}. \tag{4.38}$$

By varying the 9 coefficients over $\mathbb{R}^9$ the vector field $\mathbf{w}_h$ spans $\mathcal{W}_h$, with each set of coefficients corresponding to exactly one vector field and vice versa. The dimension of $\mathcal{W}_h$ is certainly 9. The next step is to build a basis of 9 vector fields for $\mathcal{W}_h$.

When one encounters a vector space defined by a linear expression involving a set of arbitrary real coefficients $c_1, \ldots, c_m$, one can always build a basis of the space by setting the coefficients to the values of the canonical basis of $\mathbb{R}^m$, that is,

$$(1,0,\ldots,0)^T, \quad (0,1,0,\ldots,0)^T, \quad \ldots$$

This leads to the following basis of $\mathcal{V}_h$ (we adopt $\mathbf{N}_k$ as notation for the elements of the basis to emphasize that they are vector fields over Ω):

$$\mathbf{N}_1(x_1,x_2) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \mathbf{N}_2(x_1,x_2) = \begin{pmatrix} x_1 \\ 0 \end{pmatrix}, \qquad \mathbf{N}_3(x_1,x_2) = \begin{pmatrix} x_2 \\ 0 \end{pmatrix},$$

$$\mathbf{N}_4(x_1,x_2) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \mathbf{N}_5(x_1,x_2) = \begin{pmatrix} 0 \\ x_1 \end{pmatrix}, \qquad \mathbf{N}_6(x_1,x_2) = \begin{pmatrix} 0 \\ x_2 \end{pmatrix},$$

$$\mathbf{N}_7(x_1,x_2) = \begin{pmatrix} 0 \\ x_1^2 \end{pmatrix}, \qquad \mathbf{N}_8(x_1,x_2) = \begin{pmatrix} 0 \\ x_1 x_2 \end{pmatrix}, \qquad \mathbf{N}_9(x_1,x_2) = \begin{pmatrix} 0 \\ x_2^2 \end{pmatrix}.$$

Remark: One could use any other basis of $\mathbb{R}^m$, which would lead to a different basis of $\mathcal{W}_h$.

The choice of the basis is not irrelevant. Remember that the space $\mathcal{V}_h$ is defined, as usual, as

$$\mathcal{V}_h = \{\mathbf{v}_h \in \mathcal{W}_h \mid v_{h1}(x_1 = 0, x_2) = 0,\ v_{h2}(x_1 = 0, x_2) = 0\}\ ,$$

and we need to take a basis of $\mathcal{W}_h$ **of which a subset is a basis of** $\mathcal{V}_h$.

Fortunately, our judicious choice of coordinates and location of the origin makes the basis $\mathbf{N}_1$ above an adequate one. The only fields in $\mathcal{W}_h$ that are zero at the left boundary are those belonging to

$$\mathcal{V}_h = \left\{\mathbf{v}_h \in \mathcal{W}_h \mid \mathbf{v}_h = \left(c_2 x_1\ ,\ c_5 x_1 + c_7 x_1^2 + c_8 x_1 x_2\right)^T\right\}\ . \tag{4.39}$$

Thus, the functions $\mathbf{N}_2$, $\mathbf{N}_5$, $\mathbf{N}_7$ and $\mathbf{N}_8$ are a basis of $\mathcal{V}_h$, **which in this case coincides with** $\mathcal{S}_h$ because the Dirichlet conditions are all zero.

Once we have established a suitable basis, the next step is as usual to build the stiffness matrix entries $\mathsf{K}_{ij} = a(\mathbf{N}_j, \mathbf{N}_i)$ and the load vector entries $\mathsf{f}_i = \ell(\mathbf{N}_i)$, for $i \in \eta \setminus \eta_g = \{2,5,7,8\}$. The entries with $i \in \eta_g$ will be taken from the identity matrix and the imposed boundary condition, so as to enforce, in this specific case,

$$c_1 = c_3 = c_4 = c_6 = c_9 = 0\ .$$

The computation of the stiffness matrix is easier to understand if we introduce a notation for the strain, divergence and stress fields induced by each basis function, i.e.,

$$B^i = \varepsilon(\nabla \mathbf{N}_i), \qquad D^i = \operatorname{div} \mathbf{N}_i, \qquad \Sigma^i = \sigma(\nabla \mathbf{N}_i)\ . \tag{4.40}$$

Explicitly,

$$B^1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad B^2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B^3 = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix},$$

$$B^4 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad B^5 = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}, \quad B^6 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

$$B^7 = \begin{pmatrix} 0 & x_1 \\ x_1 & 0 \end{pmatrix}, \quad B^8 = \begin{pmatrix} 0 & \frac{x_2}{2} \\ \frac{x_2}{2} & x_1 \end{pmatrix}, \quad B^9 = \begin{pmatrix} 0 & 0 \\ 0 & 2x_2 \end{pmatrix}.$$

The trace of B^i equals D^i, so

$$D^1 = 0, \quad D^2 = 1, \quad D^3 = 0, \quad D^4 = 0, \quad D^5 = 0,$$

$$D^6 = 1, \quad D^7 = 0, \quad D^8 = x_1, \quad D^9 = 2x_2.$$

The matrix expression for Σ^i is

$$\Sigma^i = \underbrace{\frac{E}{1+\nu}}_{2\mu} B^i + \underbrace{\frac{E\nu}{(1+\nu)(1-2\nu)}}_{\lambda} D^i \mathbf{I}\,,$$

where we have introduced the Lamé coefficients μ and λ. We thus have

$$\Sigma^1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma^2 = \begin{pmatrix} 2\mu+\lambda & 0 \\ 0 & \lambda \end{pmatrix}, \quad \Sigma^3 = \begin{pmatrix} 0 & \mu \\ \mu & 0 \end{pmatrix},$$

$$\Sigma^4 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma^5 = \begin{pmatrix} 0 & \mu \\ \mu & 0 \end{pmatrix}, \quad \Sigma^6 = \begin{pmatrix} 0 & 0 \\ 0 & 2\mu+\lambda \end{pmatrix},$$

$$\Sigma^7 = \begin{pmatrix} 0 & 2\mu x_1 \\ 2\mu x_1 & 0 \end{pmatrix}, \quad \Sigma^8 = \begin{pmatrix} \lambda x_1 & \mu x_2 \\ \mu x_2 & (2\mu+\lambda)x_1 \end{pmatrix}, \quad \Sigma^9 = \begin{pmatrix} 2\lambda x_2 & 0 \\ 0 & (4\mu+2\lambda)x_2 \end{pmatrix}.$$

In terms of these matrices, it holds that

$$a(\mathbf{N}_j, \mathbf{N}_i) = \int_\Omega \Sigma^j : B^i \, d\Omega. \tag{4.41}$$

These integrals are easy to calculate, for example

$$\begin{aligned} \mathsf{K}_{87} &= a(\mathbf{N}_7, \mathbf{N}_8) \\ &= \int_0^L \int_{-\frac{H}{2}}^{\frac{H}{2}} \Sigma^7 : B^8 \, dx_1 dx_2 \\ &= \int_0^L \int_{-\frac{H}{2}}^{\frac{H}{2}} 2\mu x_1 x_2 \, dx_1 dx_2 \\ &= 0, \end{aligned}$$

or

$$\begin{aligned} \mathsf{K}_{77} &= a(\mathbf{N}_7, \mathbf{N}_7) \\ &= \int_0^L \int_{-\frac{H}{2}}^{\frac{H}{2}} \Sigma^7 : B^7 \, dx_1 dx_2 \\ &= \int_0^L \int_{-\frac{H}{2}}^{\frac{H}{2}} 4\mu x_1^2 \, dx_1 dx_2 \\ &= \frac{4\mu H L^3}{3} . \end{aligned}$$

After computing all necessary integrals we arrive at the stiffness matrix

$$\mathsf{K} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & (2\mu+\lambda)HL & 0 & 0 & 0 & 0 & 0 & \frac{\lambda HL^2}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu HL & 0 & \mu HL & 0 & \mu HL^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \mu HL^2 & 0 & \mu HL^2 & 0 & \frac{4\mu HL^3}{3} & 0 & 0 \\ 0 & \frac{\lambda HL^2}{2} & 0 & 0 & 0 & \frac{(2\mu+\lambda)HL^2}{2} & 0 & \frac{\mu H^3 L}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} .$$

The load vector F has

$$\mathsf{F}_1 = \mathsf{F}_3 = \mathsf{F}_4 = \mathsf{F}_6 = \mathsf{F}_9 = 0$$

because the corresponding coefficients are imposed as zero. For $i \in \{2,5,7,8\}$, we have to compute

$$\mathsf{F}_i = \ell(\mathbf{N}_i) = -\int_\Omega \rho \, g \, N_2^i(x_1, x_2) \, dx_1 dx_2$$

so,

$$\begin{aligned}
\mathsf{F}_2 &= -\int_0^L \int_{-\frac{H}{2}}^{\frac{H}{2}} \rho\, g\, 0 \, dx_1 dx_2 = 0 \\
\mathsf{F}_5 &= -\int_0^L \int_{-\frac{H}{2}}^{\frac{H}{2}} \rho\, g\, x_1 \, dx_1 dx_2 = -\frac{\rho g H L^2}{2} \\
\mathsf{F}_7 &= -\int_0^L \int_{-\frac{H}{2}}^{\frac{H}{2}} \rho\, g\, x_1^2 \, dx_1 dx_2 = -\frac{\rho g H L^3}{3} \\
\mathsf{F}_8 &= -\int_0^L \int_{-\frac{H}{2}}^{\frac{H}{2}} \rho\, g\, x_1 x_2 \, dx_1 dx_2 = 0
\end{aligned}$$

We are now in a position to solve the linear system

$$\mathsf{K}\,\mathsf{U} = \mathsf{F}\,,$$

to obtain the unknown vector of coefficients

$$\mathsf{U} = (c_1, c_2, \ldots, c_9)^T\,.$$

We adopt, as an example, the values $H = 0.1$ m, $L = 1$ m, $E = 2 \times 10^{11}$ Pa, $\nu = 0.3$, $\rho = 8000\,\mathrm{kg/m^3}$, $g = 9.8\,\mathrm{m/s^2}$. The corresponding Lamé coefficients are $\mu = 7.692 \times 10^{10}$ Pa and $\lambda = 1.154 \times 10^{11}$ Pa.

The solution of the linear system is

$$\mathsf{U} = \left(\,0\,,\,0\,,\,0\,,\,0\,,\,-1.0192\times 10^{-6}\,,\,0\,,\,0.5096\times 10^{-6}\,,\,0\,,\,0\,\right)^T\,.$$

Equivalently, the unknown coefficients turn out to be

$$c_2 = 0, \qquad c_5 = -1.0192\times 10^{-6}, \qquad c_7 = 0.5096\times 10^{-6}\mathrm{m}^{-1}, \qquad c_9 = 0.$$

This means that the numerical approximation to the displacement field is given by

$$\mathbf{u}_h(x_1, x_2) = \begin{pmatrix} 0 \\ -1.0192\times 10^{-6}\times x_1 + 0.5096\times 10^{-6}\mathrm{m}^{-1}\times x_1^2 \end{pmatrix}.$$

The deformed geometry in which each point is moved according to

$$\mathbf{x} \mapsto \mathbf{x} + \alpha\, \mathbf{u}_h(\mathbf{x}),$$

where $\alpha = 10^5$ is a scaling factor added to render the deformation visible, is shown in Fig. 4.4. Notice that the maximum displacement occurs at $x_1 = L$ (which is logical), and the value is

$$\mathbf{u}_h(x_1, x_2 = L) = \begin{pmatrix} 0 \\ -0.5096\,\mathrm{microns} \end{pmatrix}.$$

The associated strain and stress fields are

$$\epsilon(\mathbf{u}_h) = \sum_{j=1}^{9} c_j\, \epsilon(\mathbf{N}_j) = \sum_{j=1}^{9} c_j\, B^j = \begin{pmatrix} 0 & \frac{1}{2}c_5 + c_7\, x_1 \\ \text{symm} & 0 \end{pmatrix}$$

$$\sigma(\mathbf{u}_h) = \sum_{j=1}^{9} c_j\, \sigma(\mathbf{N}_j) = \sum_{j=1}^{9} c_j\, \Sigma^j = \begin{pmatrix} 0 & \left(\frac{1}{2}c_5 + c_7\, x_1\right)\mu \\ \text{symm} & 0 \end{pmatrix}$$

The maximum shear strain and stress occurs at $x_1 = 0$. The values are 0.5096×10^{-6} and 39200 Pa.

A salient feature of the approximate solution obtained with the basis functions $\{\mathbf{N}_1, \ldots, \mathbf{N}_9\}$ is that it is very wrong.

The space $\mathscr{S}_h$, which is given by (4.39), keeps the vertical planes vertical (because v_{h1} is independent of x_2). This space is not suitable for representing the typical bending deformations, characterized by the rotation of the vertical planes with the (small) angle varying with x_1.

To improve the result one needs to enlarge the space $\mathscr{S}_h$, which is accomplished by enlarging the encompassed space $\mathscr{W}_h$. More specifically, **let us add three aditional basis functions $\mathbf{N}_{10}$, $\mathbf{N}_{11}$ and $\mathbf{N}_{12}$** so as to complete the quadratic polynomials in the horizontal component:

$$\mathbf{N}_{10}(x_1, x_2) = \begin{pmatrix} x_1^2 \\ 0 \end{pmatrix}, \quad \mathbf{N}_{11}(x_1, x_2) = \begin{pmatrix} x_1 x_2 \\ 0 \end{pmatrix}, \quad \mathbf{N}_{12}(x_1, x_2) = \begin{pmatrix} x_2^2 \\ 0 \end{pmatrix}.$$

The crucial one, in view of what was discussed above, is $\mathbf{N}_{11}$. From them we can compute the associated strain and stress fields,

$$B^{10} = \begin{pmatrix} 2x_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B^{11} = \begin{pmatrix} x_2 & \frac{x_1}{2} \\ \frac{x_1}{2} & 0 \end{pmatrix}, \quad B^{12} = \begin{pmatrix} 0 & x_2 \\ x_2 & 0 \end{pmatrix}.$$

$$D^{10} = 2x_1, \quad D^{11} = x_2, \quad D^{12} = 0.$$

$$\Sigma^{10} = \begin{pmatrix} (4\mu + 2\lambda)x_1 & 0 \\ 0 & 2\lambda x_1 \end{pmatrix}, \quad \Sigma^{11} = \begin{pmatrix} (2\mu + \lambda)x_2 & \mu x_1 \\ \mu x_1 & \lambda x_2 \end{pmatrix},$$

$$\Sigma^{12} = \begin{pmatrix} 0 & 2\mu x_2 \\ 2\mu x_2 & 0 \end{pmatrix}.$$

From this it is straightforward to enlarge the previous 9×9 stiffness matrix into the 12×12 new one. Notice that c_{12} **must be zero** to satisfy the boundary conditions. The new 12×1 load vector array is just the previous one with 3 zeros added at the end (why?).

The algebraic solution obtained with 12 basis functions ($k_1 = k_2 = 2$) is

$$\mathsf{U} = 10^{-6} \times (\, 0\,,\, 0\,,\, 0\,,\, 0\,,\, -1.0192\,,\, 0\,,\, -28.61\,,\, 0\,,\, 0\,,\, 0\,,\, 58.24\,,\, 0\,)^T\ .$$

With respect to the previous solution, one observes a significant change in c_7 while c_5 remains the same. The c_{11} coefficient, which did not exist previously, takes a relatively large value, while c_{10} is zero. A deformation mode that combines the functions $\mathbf{N}_7$ and $\mathbf{N}_{11}$ is activated by the load (this mode was not possible when $\mathbf{N}_{11}$ was not there).

The new variational method approximation to $\mathbf{u}$ is

$$\mathbf{u}_h(x_1, x_2) = \begin{pmatrix} 58.24 \times 10^{-6}\mathrm{m}^{-1} \times x_1 \times x_2 \\ -1.0192 \times 10^{-6} \times x_1 - 28.61 \times 10^{-6}\mathrm{m}^{-1} \times x_1^2 \end{pmatrix}.$$

It is shown in Fig. 4.5, but notice that the scaling factor α has been reduced to 10^4 because the displacements are much larger. The solution now makes physical sense. A typical bending deformation takes place. The vertical displacement of the tip of the bar is -29.63 microns (almost 60 times larger in magnitude than with the previous space). The computation of the corresponding strain and stress fields is straightforward.

Food for thought: The solution to the variational method only "uses" the basis functions $\mathbf{N}_5$, $\mathbf{N}_7$ and $\mathbf{N}_{11}$. If we had chosen simply $\mathcal{S}_h = \mathcal{V}_h =$ span $\{\mathbf{N}_5, \mathbf{N}_7, \mathbf{N}_{11}\}$ (with just 3 unknown coefficients!), **would we have obtained the same $\mathbf{u}_h$?** The answer is **yes** (why?).

Summary of this long example: This example illustrated the application of a variational method in elasticity problems with spaces of global polynomials. This is only possible for some domains and boundary conditions. We followed the same procedure as in Example 2.7, with the necessary adaptations brought by the vector character of the unknown $\mathbf{u}$.

During the exposition we showed that the solution of the variational method may be very inaccurate if the selected space is unable to approximate the dominant deformation modes of the exact solution.

Invertibility of the Stiffness Matrix. If the space $\mathcal{V}_h$ is made of continuous functions, then $\mathcal{V}_h \subset \mathcal{V}$, and we have that (4.35) is in particular satisfied for all $\mathbf{v}_h \in \mathcal{V}_h$ and thus $a(\cdot,\cdot)$ **is coercive on** $\mathcal{V}_h$. In other words, (3.6) in Céa's Lemma (Theorem 3.1) is satisfied. As with scalars problems, this implies that **the stiffness matrix is invertible**.

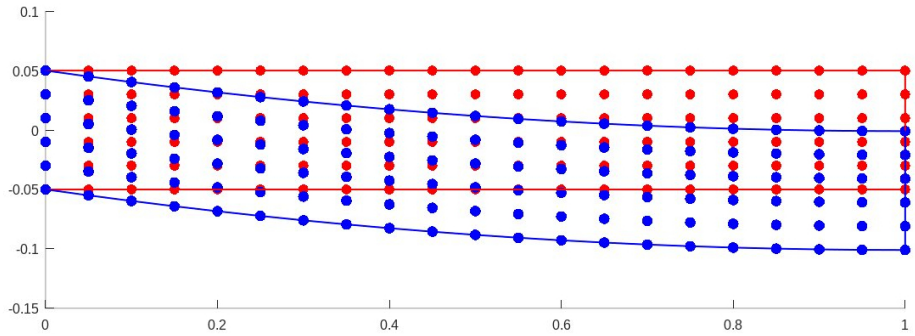


Figure 4.4 Deformed geometry (with the displacement scaled by a factor of 10^5 to render it visible) obtained with the variational method when using $k_1 = 1$ and $k_2 = 2$ (9 basis functions). Shown are the reference and deformed positions of a set of sampling points (in red and blue, respectively). Also shown are the boundaries of the reference and deformed configurations. This solution is physically unrealistic.

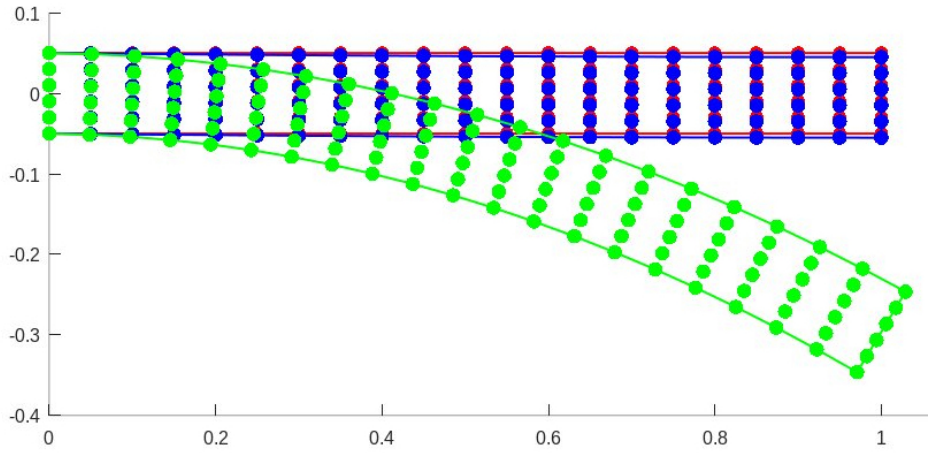


Figure 4.5 In green we show the deformed geometry (with the displacement scaled by a factor of 10^4 to render it visible) obtained with the variational method when using $k_1 = k_2 = 2$ (12 basis functions). In red and blue we show the same reference and approximate deformation as in Fig. 4.4, but now they are almost superposed because α has been decreased to 10^4.

4.4 Finite Element Spaces for Multifield problems in 2D

In Chapter 2 we introduced the P_1 finite element space in two space dimensions. The P_1 finite element provides basis functions that are **continuous scalar functions** which, by refining the mesh, can approximate any function $u \in C^1(\Omega)$.

Though they have not yet been discussed in this notes, there exist P_k finite elements that provide continuous scalar functions which are piecewise polynomials of any order $k \geq 1$ in the variables $x_1 - x_2$.

None of these finite element spaces can *per se* approximate problems with several unknowns such as the 2D elasticity problem, in which the unknown is a vector field **u** which, upon selecting a suitable coordinate frame, becomes **a pair** of scalar functions (u_1, u_2). How to build a finite element space that can approximate **u**? This is an example of one among many **multifield** problems in Engineering.

Multifield problems. These are problems in which there exist $p > 1$ coupled unknowns, $u_1, \ldots, u_p$. Typical examples are thermoelasticity (temperature and displacement), thermal convection (temperature and fluid velocity and pressure), electrodynamics (electric and magnetic fields), etc. The solution is thus a p-tuple $(u_1, u_2, \ldots, u_p)$, which will be organized as a column array for convenience.

We thus pose the general question of **how to build multifields using scalar finite elements as raw material**. The following lemma describes a procedure to do that for **a two-field in two-dimensional domain**. The extension to p-fields or to other space dimensions is straightforward.

Lemma 4.2. *Let $\mathcal{W}_{h1}$ and $\mathcal{W}_{h2}$ be two finite element spaces with bases $\mathcal{N}^\Omega = \{N_1, N_2, \ldots, N_{m_1}\}$ and $\mathcal{M}^\Omega = \{M_1, M_2, \ldots, M_{m_2}\}$. Then,*

a) *the space of two-fields*

$$\mathcal{W}_h = \left\{ \underline{w_h}(x_1, x_2) = \begin{pmatrix} w_{h1}(x_1, x_2) \\ w_{h2}(x_1, x_2) \end{pmatrix} \mid w_{h1} \in \mathcal{W}_{h1},\ w_{h2} \in \mathcal{W}_{h2} \right\} = \mathcal{W}_{h1} \times \mathcal{W}_{h2} \tag{4.42}$$

is a vector space of dimension $m_1 + m_2$.

b) *The set of two-fields*

$$\mathcal{N}^\Omega = \left\{ \underbrace{\begin{pmatrix} N_1 \\ \\ 0 \end{pmatrix}}_{\underline{N}_1}, \underbrace{\begin{pmatrix} N_2 \\ \\ 0 \end{pmatrix}}_{\underline{N}_2}, \ldots \underbrace{\begin{pmatrix} N_{m_1} \\ \\ 0 \end{pmatrix}}_{\underline{N}_{m_1}}, \underbrace{\begin{pmatrix} 0 \\ \\ M_1 \end{pmatrix}}_{\underline{N}_{m_1+1}}, \underbrace{\begin{pmatrix} 0 \\ \\ M_2 \end{pmatrix}}_{\underline{N}_{m_1+2}}, \ldots \underbrace{\begin{pmatrix} 0 \\ \\ M_{m_2} \end{pmatrix}}_{\underline{N}_{m_1+m_2}} \right\}$$

is a basis of $\mathcal{W}_h$.

c) *Assume that $\mathcal{W}_{h1}$ and $\mathcal{W}_{h2}$ are generated, respectively, by* **scalar finite elements** *$(\Omega_e, \mathcal{N}^e)$ and $(\Omega_e, \mathcal{M}^e)$* **on the same partition** *$\Omega = \cup_e \Omega_e$, where*

$$\mathcal{N}^e = \{N_1^e, N_2^e, \ldots, N_{\ell_1}^e\}, \qquad \mathcal{M}^e = \{M_1^e, M_2^e, \ldots, M_{\ell_2}^e\}$$

are the corresponding sets of **scalar shape functions**. *Then* $(\Omega_e, \mathcal{N}^e)$ *is a* **two-field finite element**, *where*

$$\mathcal{N}^e = \Big\{ \underbrace{\begin{pmatrix} N_1^e \\ \\ 0 \end{pmatrix}}_{\underline{N}_1^e}, \underbrace{\begin{pmatrix} N_2^e \\ \\ 0 \end{pmatrix}}_{\underline{N}_2^e}, \ldots \underbrace{\begin{pmatrix} N_{\ell_1}^e \\ \\ 0 \end{pmatrix}}_{\underline{N}_{\ell_1}^e}, \underbrace{\begin{pmatrix} 0 \\ \\ M_1^e \end{pmatrix}}_{\underline{N}_{\ell_1+1}^e}, \underbrace{\begin{pmatrix} 0 \\ \\ M_2^e \end{pmatrix}}_{\underline{N}_{\ell_1+2}^e}, \ldots \underbrace{\begin{pmatrix} 0 \\ \\ M_{\ell_2}^e \end{pmatrix}}_{\underline{N}_{\ell_1+\ell_2}^e} \Big\}$$

d) *Let* LG1 *and* LG2 *be the* **local-to-global maps** *that establish the correspondences between* $\mathcal{N}^e$ *and* $\mathcal{N}^\Omega$ *and between* $\mathcal{M}^e$ *and* $\mathcal{M}^\Omega$, *respectively; i.e.,*

$$N_A = \mathop{\mathring{\sum}}_{\{(a,e)|\mathsf{LG1}(a,e)=A\}} N_a^e, \qquad M_B = \mathop{\mathring{\sum}}_{\{(b,e)|\mathsf{LG2}(b,e)=B\}} M_b^e \,. \tag{4.43}$$

Then the **local-to-global map** LG **of the two-field finite element** *that satisfies*

$$\underline{N}_A = \mathop{\mathring{\sum}}_{\{(a,e)|\mathsf{LG}(a,e)=A\}} \underline{N}_a^e \tag{4.44}$$

is given by

$$\mathsf{LG}(a,e) = \begin{cases} \mathsf{LG1}(a,e) & \text{if } 1 \le a \le \ell_1 \\ \mathsf{LG2}(a-\ell_1,e) + m_1 & \text{if } \ell_1 + 1 \le a \le \ell_1 + \ell_2 \end{cases}$$

In Octave/MATLAB the array LG *can be formed with the instruction*

```
LG = [ LG1 ; LG2 + m1 ];
```

e) *Assume that both spaces* $\mathcal{W}_{h1}$ *and* $\mathcal{W}_{h2}$ *satisfy the approximability property, i.e., assume that for any smooth two-field* $\underline{w}$ *there exist constants* C_1, C_2, k_1 *and* k_2 *and norms* $\|\cdot\|_{W_1}$ *and* $\|\cdot\|_{W_2}$, *all independent of* h, *such that*

$$\min_{v_h \in \mathcal{W}_{h1}} \|w_1 - v_h\|_{W_1} \le C_1 h^{k_1} \quad \text{and} \quad \min_{v_h \in \mathcal{W}_{h2}} \|w_2 - v_h\|_{W_2} \le C_2 h^{k_2} \,. \tag{4.45}$$

Then,

$$\min_{\underline{w_h} \in \mathcal{W}_h} \left(\|w_1 - w_{h1}\|_{W_1}^2 + \|w_2 - w_{h2}\|_{W_2}^2 \right)^{\frac{1}{2}} \le C_1 h^{k_1} + C_2 h^{k_2} \,. \tag{4.46}$$

This lemma allows us to exploit the scalar finite element spaces we have introduced in previous chapters to build multifield finite element spaces. We do this with the P_1-space in two dimensions in the following section.

4.5 Solving Linear Elasticity Problems in 2D with P_1 Finite Elements

In 2D Cartesian coordinates the displacement field **u** is represented by a two-field $\underline{u} = (u_1, u_2)^T$. Also, the space $\mathscr{W} = \mathbf{H}^1(\Omega)$ of linear elasticity coincides, by its definition (4.25), with $\mathscr{W}_1 \times \mathscr{W}_2$ where $\mathscr{W}_1 = \mathscr{W}_2 = H^1(\Omega)$.

We can thus identify the two-fields with the vector fields and keep using the boldface vector notation that was adopted from the beginning of this chapter.

The **two-field finite element space** that we adopt here is a very popular one. Each component of the field is approximated with P_1 finite elements based on a triangulation of the domain.

Let $\mathscr{T}_h$ be a **conforming triangulation of the domain**, with its corresponding arrays X, LV. Let W_h be the **scalar P_1-space associated to $\mathscr{T}_h$**.

Then, the P_1 **space for 2D elasticity** is simply

$$\mathscr{W}_h = W_h \times W_h = \left\{ \mathbf{w}_h = (w_{h1}, w_{h2})^T : \Omega \to \mathbb{R}^2 \mid w_{h1} \in W_h,\ w_{h2} \in W_h \right\} . \qquad (4.47)$$

In this situation, we can apply Lemma 4.2 to build the basis and the local-to-global map. In particular,

```
nod = size(X,2);
LG = [ LV ; LV+nod ];
```

Let us now turn to the element stiffness matrix and element load vector.

4.5.1 Element stiffness matrix and element load vector

The element basis functions are, in terms of the scalar basis functions N_1^e, N_2^e and N_3^e that we already know well (the barycentric coordinates λ_1, λ_2 and λ_3),

$$\mathbf{N}_1^e = \begin{pmatrix} N_1^e \\ 0 \end{pmatrix}, \quad \mathbf{N}_2^e = \begin{pmatrix} N_2^e \\ 0 \end{pmatrix}, \quad \mathbf{N}_3^e = \begin{pmatrix} N_3^e \\ 0 \end{pmatrix},$$

$$\mathbf{N}_4^e = \begin{pmatrix} 0 \\ N_1^e \end{pmatrix}, \quad \mathbf{N}_5^e = \begin{pmatrix} 0 \\ N_2^e \end{pmatrix}, \quad \mathbf{N}_6^e = \begin{pmatrix} 0 \\ N_3^e \end{pmatrix}.$$

We will use the same array

$$\mathsf{dN}_{ib} = \frac{\partial N_b^e}{\partial x_i}$$

as in Chapter 2. From dN we can compute the strain fields $B^a = \varepsilon(\nabla \mathbf{N}_a)$.

$$B^1 = \begin{pmatrix} \mathsf{dN}_{11} & \frac{1}{2}\mathsf{dN}_{21} \\ \frac{1}{2}\mathsf{dN}_{21} & 0 \end{pmatrix}, B^2 = \begin{pmatrix} \mathsf{dN}_{12} & \frac{1}{2}\mathsf{dN}_{22} \\ \frac{1}{2}\mathsf{dN}_{22} & 0 \end{pmatrix}, B^3 = \begin{pmatrix} \mathsf{dN}_{13} & \frac{1}{2}\mathsf{dN}_{23} \\ \frac{1}{2}\mathsf{dN}_{23} & 0 \end{pmatrix},$$

$$B^4 = \begin{pmatrix} 0 & \frac{1}{2}\mathsf{dN}_{11} \\ \frac{1}{2}\mathsf{dN}_{11} & \mathsf{dN}_{21} \end{pmatrix}, B^5 = \begin{pmatrix} 0 & \frac{1}{2}\mathsf{dN}_{12} \\ \frac{1}{2}\mathsf{dN}_{12} & \mathsf{dN}_{22} \end{pmatrix}, B^6 = \begin{pmatrix} 0 & \frac{1}{2}\mathsf{dN}_{13} \\ \frac{1}{2}\mathsf{dN}_{13} & \mathsf{dN}_{23} \end{pmatrix}.$$

The way to store three-dimensional arrays of matrices in Octave/MATLAB is by adding an additional index **as last index**. The array of strain matrices can thus be coded as

```
for i=1:3
   BB(1:2,1:2,i)  = [dN(1,i),0.5*dN(2,i); 0.5*dN(2,i),0];
   BB(1:2,1:2,i+3)= [0,0.5*dN(1,i); 0.5*dN(1,i),dN(2,i)];
end
```

and the stresses $\Sigma^a = \sigma(\nabla \mathbf{N}^a)$ (again using the Lamé coefficients μ and λ, which are assumed constant in the element with value μ_e and λ_e, respectively) as,

```
for a=1:6
   DD(a)=sum(diag(BB(:,:,a));
   SS(:,:,a)=2*mue*BB(:,:,a)+lambdae*DD(a)*eye(2);
end
```

where `eye(2)` is the two-dimensional identity matrix.

Remark: It is also possible (and convenient) to store these symmetric matrices as one-dimensional arrays by using the **Voigt notation**.

The **element stiffness matrix** K^e is given by

$$\mathsf{K}^e_{ab} = a(\mathbf{N}^b, \mathbf{N}^a) = \int_{\Omega_e} \Sigma^b : B^a \, dx_1 dx_2 \; = A_e \Sigma^b : B^a \,, \tag{4.48}$$

where we have used that **both B^a and Σ^b are constant in the element**.

Remark: The P_1 element in elasticity is also called CST element, for Constant Strain Triangle.

The **element load vector** is also easy to compute. Leaving the treatment of the Neumann boundary conditions for later, let us assume that the body force $\mathbf{b}$ is constant and equal to $\mathbf{b}^e$ in the element. Then,

$$\mathsf{F}^e_a = \ell(\mathbf{N}^e_a) = \int_{\Omega_e} \mathbf{b}^e \cdot \mathbf{N}^e_a \, dx_1 dx_2 = \begin{cases} \frac{A_e}{3} \, b^e_1 & \text{if } 1 \le a \le 3 \\ \frac{A_e}{3} \, b^e_2 & \text{if } 4 \le a \le 6 \end{cases}$$

The double contraction ": " of two matrices A and B can be coded in Octave/-MATLAB as

```
contraction = sum(sum(A.*B));
```

Putting it all together, the following lines compute the element stiffness matrix and element load vector.

```
function [Ke, Fe]=elementKandF(xe,Ee,nue,be)
  mue=Ee/(1+nue)/2;
  lambdae=mue*nue/(1-2*nue);
  dN=[xe(2,2)-xe(2,3),xe(2,3)-xe(2,1),xe(2,1)-xe(2,2);...
      xe(1,3)-xe(1,2),xe(1,1)-xe(1,3),xe(1,2)-xe(1,1)];
  Ae2=dN(2,3)*dN(1,2)-dN(1,3)*dN(2,2);
  dN=dN/Ae2; Ae=Ae2/2;
  %% Compute strains and stresses
  for i=1:3
```

```
    BB(1:2,1:2,i)  = [dN(1,i),0.5*dN(2,i); 0.5*dN(2,i),0];
    BB(1:2,1:2,i+3)= [0,0.5*dN(1,i); 0.5*dN(1,i),dN(2,i)];
  end
  for a=1:6
    DD(a)=sum(diag(BB(:,:,a)));
    SS(:,:,a)=2*mue*BB(:,:,a)+lambdae*DD(a)*eye(2);
  end
  %% Build Ke and Fe
  for a=1:6
   for b=1:6
    Ke(a,b)=Ae*sum(sum(SS(:,:,b).*BB(:,:,a)));
   end
  end
  Fe=Ae/3*[be(1);be(1);be(1);be(2);be(2);be(2)];
end
```

4.5.2 Boundary conditions

The Dirichlet and Neumann boundary conditions can be treated with a slight adaptation of the techniques used in Chapter 2 to make it suitable for vector fields.

The array of boundary edges BE, remember, has as many columns as boundary edges in the mesh. The first line and second lines contain the two nodes of each boundary edge. The third line contains the geometrical line to which each edge belongs.

The array η_g (EtaG) must now include both unknowns (u_{h1} and u_{h2}) of all nodes that have Dirichlet conditions. It is possible to have boundaries along which only one of the unknowns is imposed, as would be the case of a symmetry line. Since η_g imposes constraints on each degree of freedom, η_g contains all the indices of degrees of freedom whose values are prescribed by the essential boundary conditions.

Similarly, the array H has now two lines. Each column gives us the surface force **H** that is applied to each boundary edge (assumed constant each edge).

We show below the procedure to build η_g, GG and H in a case in which the Dirichlet boundary conditions (equal to zero) are applied along geometrical line #4 and a surface force with value $(0,-100)^T$ is applied along geometrical line #2. Notice that we create the scalar version on η_g first and then transform it to the vector version. This would not be correct if some nodes have the u_1 component imposed but not u_2 or viceversa.

```
bcaux=zeros(1,size(X,2));
nbe=size(BE,2);
HH=zeros(2,nbe);
for k=1:nbe
   if (BE(3,k)==4)
      bcaux(BE(1,k))=1;
      bcaux(BE(2,k))=1;
```

```
   end
   if (BE(3,k)==2)
      HH(:,k)=[0;-1000];
   end
end
EtaG=find(bcaux==1);
%% This is the EtaG for scalars, for vectors it is
EtaG=[EtaG EtaG+nod];
%% Set boundary values to zero, both components.
GG=0*ones(1,length(EtaG));
```

Let us use the same trick as in Chapter 2. We build the stiffness matrix and load vector as if there were no Dirichlet conditions or Neumann conditions, then we add the Neumann contributions, and just before solving the system we correct K and F according to η_g and GG (the list of boundary values).

The procedure can be easily coded as

```
%% Neumann contributions
for ied=1:nbe
  lged=BE(1:2,ied);
  xed(1:dd,1:2)=X(1:dd,lged);
  Led=norm(xed(:,1)-xed(:,2));
  Hed=HH(:,ied);
  F(lged)+=Hed(1)*Led/2;
  F(lged+nod)+=Hed(2)*Led/2;
end
%% unknowns with specified value
ng=length(EtaG);
II=eye(nunk);
for ig=1:ng
 K(EtaG(ig),:)=II(EtaG(ig),:);
 F(EtaG(ig))=GG(ig);
end
```

4.5.3 Example: A cantilever bar

We go back to the geometry and conditions of Fig. 4.3, in which a cantilever bar is fixed at its left side to a rigid vertical wall. We use the same coefficients as in Example 4.5, namely $H = 0.1$ m, $L = 1$ m, $E = 2 \times 10^{11}$ Pa, $\nu = 0.3$, $\rho = 8000$ kg/m^3, $g = 9.8$ m/s^2. We want to illustrate that the finite element method indeed provides a convergent approximation for this problem.

As done in Example 4.5, we will mainly visualize the deformed geometry (affected by $\alpha = 10^4$ as in Fig. 4.5) and report the vertical displacement of the upper right corner of the bar (node #3).

The polygonal boundary was defined from the points $(0, -\frac{H}{2})^T$, $(L, -\frac{H}{2})^T$, $(L, \frac{H}{2})^T$ and $(0, \frac{H}{2})^T$. The Dirichlet boundary is geometrical line #4 (between points 4 and 1). The Neumann boundary is defined as line #2 (between points 2 and 3), but **H** is taken as zero and thus has no effect.

The vertical displacement of node 3, as obtained with meshes of different refinement, is as follows

n_{nod}	n_{el}	u_{h2}(node 3)
32	40	-32.572 μm
66	86	-42.112 μm
246	402	-55.186 μm
938	1698	-59.290 μm
3530	6706	-60.221 μm
13756	26806	-60.471 μm

The corresponding deformed meshes (with displacement scaled by $\alpha = 10^4$) are shown in Fig. 4.6. As observed with the global polynomial case, the discretized bar shows to be stiffer when the mesh is coarser. The vertical displacement of node 3 converges to $\simeq -60.4\,\mu$m as the mesh is refined. In the bottom part of the figure the difference in u_{h2}(node 3) between successive meshes is plotted vs. the number of elements. The logarithmic plot shows that the pointwise value converges with order $O(h^2)$.

4.5.4 Computing the stresses

As mentioned before, one of the main objectives of solving an elastic problem is to compute the stress field.

For the P_1 finite element space, the stress is constant within each element. How to compute it? Once one has computed $\mathbf{u}_h$, it is quite simple. The function `elementStress` below does just that. It simply computes $\sigma = \sum_{a=1}^{6} U_a \Sigma^a$. Notice the similarity with `elementKandF`.

```
function sigmae=elementStress(xe,Ee,nue,Ue)
  mue=Ee/(1+nue)/2;
  lambdae=mue*nue/(1-2*nue);
  dN=[xe(2,2)-xe(2,3),xe(2,3)-xe(2,1),xe(2,1)-xe(2,2);...
      xe(1,3)-xe(1,2),xe(1,1)-xe(1,3),xe(1,2)-xe(1,1)];
  Ae2=dN(2,3)*dN(1,2)-dN(1,3)*dN(2,2);
  dN=dN/Ae2; Ae=Ae2/2;
  %% Compute strains and stresses
  for i=1:3
    BB(1:2,1:2,i)  = [dN(1,i),0.5*dN(2,i); 0.5*dN(2,i),0];
    BB(1:2,1:2,i+3)= [0,0.5*dN(1,i); 0.5*dN(1,i),dN(2,i)];
  end
```

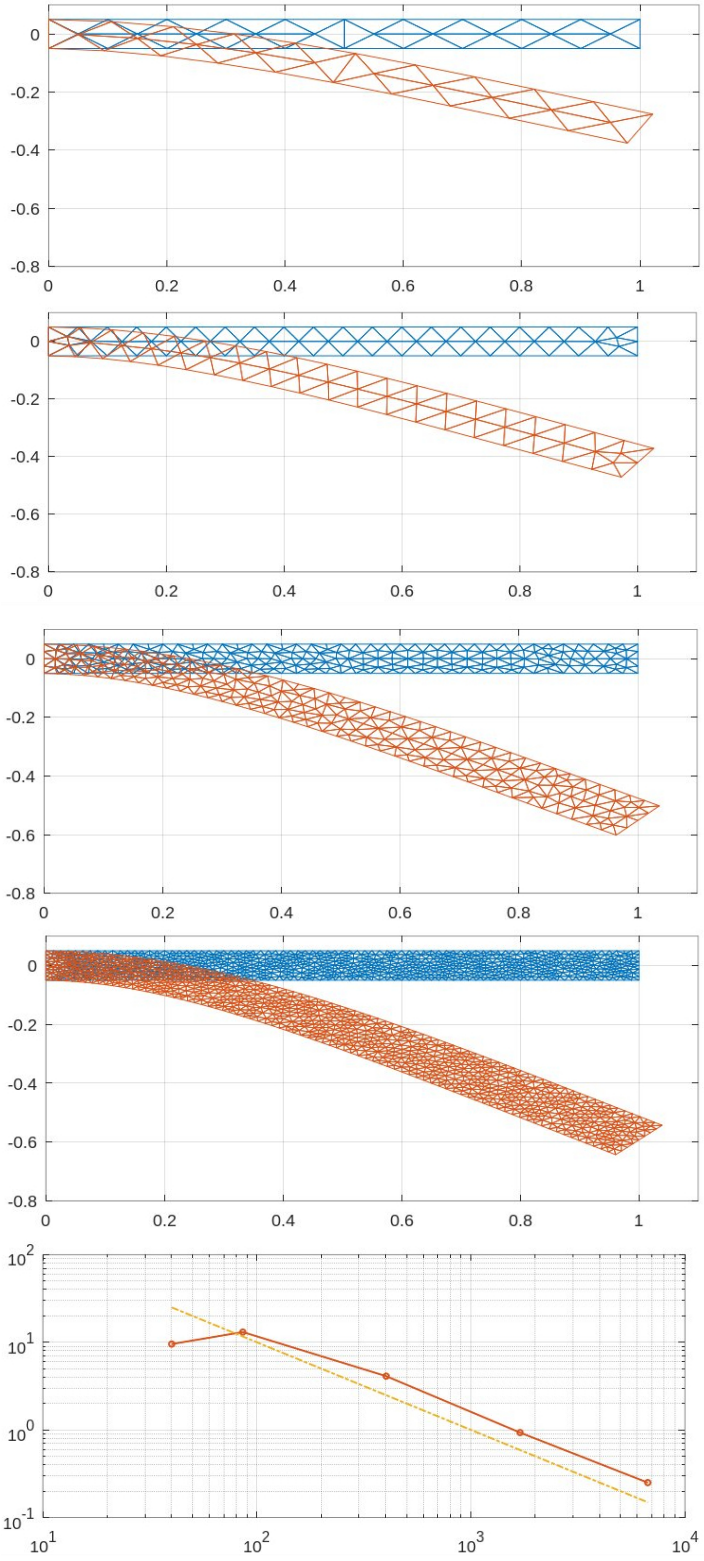


Figure 4.6 Results of a cantilever bar problem under its own weight. The first four graphs show the undeformed and deformed meshes (with the displacements scaled by $\alpha = 10^4$) for several mesh sizes. The bottom graph shows the difference (in microns) in u_{h2} at the top right corner between two successive meshes. The dotted line is a reference line with $O({n_{\text{el}}}^{-1})$ which corresponds to $O(h^2)$.

```
  for i=1:6
    DD(i)=sum(diag(BB(:,:,i)));
    SS(:,:,i)=2*mue*BB(:,:,i)+lambdae*DD(i)*eye(2);
  end
  %% Build sigmae
  sigmae=zeros(2,2);
  for i=1:6
      sigmae=sigmae+Ue(i)*SS(:,:,i);
  end
end
```

In Figure 4.7 we show the von Mises stress $\|\sigma\| = \sqrt{\sigma : \sigma}$ by plotting a circle at the centroid of each element, colored according to $\|\sigma\|$. The maximum stresses are seen to occur, as expected, in the top left and bottom left corners of the bar. The predicted value of the von Mises stress is 2.414 MPa. In the top left element, in particular, we obtain (for a mesh with 1698 elements)

$$\sigma_h = \begin{pmatrix} 2.40 & -0.176 \\ -0.176 & 0.150 \end{pmatrix} \text{ MPa},$$

showing that the main component of the stress is horizontal traction.

Since the yield stress of steel is about 250 MPa, we can be quite sure that a cantilever steel bar of such 1 m × 0.1 m will not collapse under its own weight.

Because the stress and the strain are linear in $\nabla \mathbf{u}$, we expect that they converge to the exact value more slowly (as we refine the mesh) than the displacement $\mathbf{u}$ itself. For the P_1-element we expect

$$\|\varepsilon - \varepsilon_h\|_{0,\Omega} \leq c_1 \, h$$

and

$$\|\sigma - \sigma_h\|_{0,\Omega} \leq c_2 \, h \, .$$

for some c_1 and c_2 that depend on the (smooth) exact solution. The convergence of σ_h in the **maximum norm** is a more delicate problem, since it is very sensitive to the boundary conditions and the mesh geometry. For the meshes considered for the convergence of u_2 at node 3 we obtain the results tabulated below.

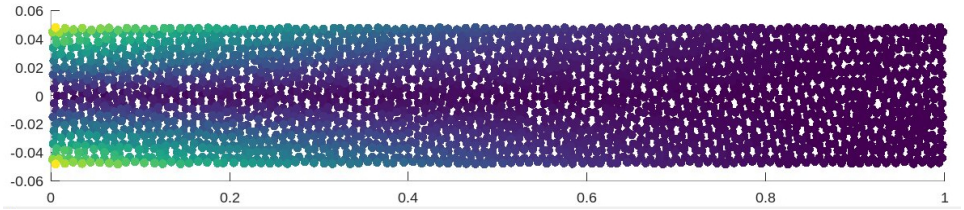


Figure 4.7 Scatter plot colored with the von Mises stress ($\|\sigma\|$). Each point is located at the centroid of an element. The maximum von Mises stress (in yellow) is 2.414 MPa and takes place at the top left and bottom left corners.

n_{nod}	n_{el}	max von Mises stress
32	40	1.098 MPa
66	86	1.802 MPa
246	402	2.103 MPa
938	1698	2.414 MPa
3530	6706	2.744 MPa
13756	26806	3.135 MPa

4.6 A Variational Method as a Minimum Principle

Let us recapitulate the steps that were followed to arrive at the discrete problem (4.37):

1. The point of departure was the Variational Problem 4.1, which introduces the potential energy function V and defines the solution $\mathbf{u}$ as the minimum of V over $\mathscr{S}$.

2. Then, using Theorem 4.1, we saw that the minimum of V over $\mathscr{S}$ satisfies the weak formulation $a(\mathbf{u},\mathbf{v}) = \ell(\mathbf{v})$ for all $\mathbf{v} \in \mathscr{V}$.

3. Once we had the weak formulation, we proceeded to the variational method of approximation just as in the previous chapters.

There is however another intuitive, even more natural, discretization method. We can go straight from item 1 above to a discrete problem by simply restricting the **minimization problem** to the discretized trial space $\mathscr{S}_h$.

In other words, one can consider the problem

Problem 4.4 (Discrete Variational Problem)**.** *Find* $\mathbf{u}_h \in \mathcal{W}_h$ *that minimizes* V *over* $\mathcal{S}_h$*; i.e., find* $\mathbf{u}_h \in \mathcal{S}_h$ *such that*

$$V(\mathbf{u}_h) \leq V(\mathbf{w}_h), \qquad \forall \mathbf{w}_h \in \mathcal{S}_h. \tag{4.49}$$

It is not difficult to see that Problem 4.4 has a unique solution, and that this solution coincides with that of the variational method. In fact, we can apply Theorem 4.1 taking as $\mathcal{W}$ the discrete space $\mathcal{W}_h$. Then $\mathcal{S}_h$ is an affine subspace of $\mathcal{W}_h$ and $\mathcal{V}_h$ is its direction, as required by the theorem. Further, notice that hypothesis (a) of the theorem is automatically satisfied, because $\mathcal{V}_h \subset \mathcal{V}$. Finally, to check that hypothesis (b) holds, notice that if **u** is the **exact solution** (assumed to exist) then

$$V(\mathbf{w}_h) \geq V(\mathbf{u}) , \qquad \forall \mathbf{w}_h \in \mathcal{S}_h .$$

This means that V is bounded from below in $\mathcal{S}_h$. In a finite-dimensional space, a function V that is bounded from below and such that $V(\mathbf{w}_h)$ tends to $+\infty$ when $\mathbf{w}_h \to \infty$ always has at least one minimum. And this is certainly the case, since the strain energy $\frac{1}{2}a(\mathbf{w}_h, \mathbf{w}_h)$ is quadratic in $\mathbf{w}_h$ and will eventually dominate and tend to $+\infty$ as $\mathbf{w}_h \to \infty$. Then V has a minimum in $\mathcal{S}_h$ as required.

Theorem 4.1 then guarantees that the minimum $\mathbf{u}_h$ of V over $\mathcal{S}_h$ is unique, and is further characterized as the unique element in $\mathcal{S}_h$ that satisfies

$$a(\mathbf{u}_h, \mathbf{v}_h) = \ell(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathcal{V}_h.$$

That is, $\mathbf{u}_h$ is the solution of the variational method.

Therefore, we have two ways of getting to $\mathbf{u}_h$: either through a variational method, or through minimization of V in $\mathcal{S}_h$. In terms of code, it tells us that we could implement our variational method solver in a totally different way. Instead of a stiffness matrix K and a load vector F, we would just need to code a *potential energy function* V, say for example

```
function V = PotentialEnergy(Y)
```

which would have as variables the unknown coefficients U_i for $i \notin \eta_g$, collected in the array Y. This function can be computed element by element in a procedure similar to the assembly procedure.

Then, instead of calling the linear system solver "\" we would call a minimization routine (such as `sqp` in Octave). The solution would be exactly the same, but the conceptual approach is quite different.

Remark: The minimization approach makes some generalizations straightforward. One could for example incorporate some restrictions to the minimization problem which would be difficult, or at least not intuitive, to formulate in both the strong or weak forms. Consider an elastic solid that is placed on top of a rigid horizontal surface. The points that are in contact with the surface are then constrained to have non-negative vertical displacement, i.e., $u_2(\mathbf{x}) \geq 0$ if $x_2 = 0$. This is a constraint that is easily incorporated into the discrete minimization

problem (not more than ten lines of code!) and leads to the correct mechanical solution. To arrive at the same numerical approximation starting from the differential equation or the weak formulation is quite involved.

We should point out here that minimization software is much more efficient if you provide the Gradient and the Hessian of the function to be minimized. To compute the gradient one needs F, and the Hessian is nothing but K, so there is no real coding difference when efficiency is important.

The variational (or minimization) approach that we have illustrated for linear elastic problems is possible whenever the bilinear form in the weak formulation is symmetric and coercive. This observation is interesting enough to devote a small section to it.

4.7 Minimization problems and variational method

Let us organize our ideas in the form of a theorem.

Theorem 4.2. *Let $\mathcal{W}$ be a normed vector space, $\mathcal{S} \subset \mathcal{W}$ an affine subspace and $\mathcal{V}$ the direction of $\mathcal{S}$. Further, let $\mathcal{W}_h \subset W$ be a finite-dimensional vector subspace, $\mathcal{S}_h = \mathcal{S} \cap \mathcal{W}_h$ and $\mathcal{V}_h = \mathcal{V} \cap \mathcal{W}_h$.*

Consider a **symmetric** *continuous bilinear form $a(\cdot,\cdot)$ defined in $\mathcal{W}$,* **coercive** *on $\mathcal{V}$, and a continuous linear form $\ell(\cdot)$, also defined in $\mathcal{W}$.*

First part: *Define the following problems:*

PA: *Find $u \in \mathcal{S}$, $V(u) < V(w)$ for all $w \in \mathcal{S}$, $w \neq u$, where*

$$V(w) = \frac{1}{2} a(w,w) - \ell(w) \,. \tag{4.50}$$

PB: *Find $u \in \mathcal{S}$ such that*

$$a(u,v) = \ell(v) \,, \qquad \forall\, v \in \mathcal{V} \,. \tag{4.51}$$

Then problems PA and PB are equivalent. *They both have unique, coincident solutions.*

Second part: *Define the following* **discrete** *problems:*

PA$_h$: *Find $u_h \in \mathcal{S}_h$, $V(u_h) < V(w_h)$ for all $w_h \in \mathcal{S}_h$, $w_h \neq u_h$.*

PB$_h$: *Find $u_h \in \mathcal{S}_h$ such that*

$$a(u_h, v_h) = \ell(v_h) \,, \qquad \forall\, v_h \in \mathcal{V}_h \,. \tag{4.52}$$

Then problems PA$_h$ and PB$_h$ are equivalent. *They both have unique, coincident solutions.*

Now let us apply this theorem to different from that of linear elasticity, in a suite of examples.

Example 4.6 (Minimization form of 1D second-order symmetric problems)

Consider, as in Sections 1 and 2 of Chapter 1, that u is the solution of the differential equation

$$-(k(x)u'(x))' + c(x)u(x) = f(x)$$

with boundary conditions $u(0) = g_0$ and $u'(L) = d_L$, where the coefficient functions k, c and f are continuous and bounded, with $k(x) \geq k_0 > 0$ and $c(x) \geq 0$ for all x.

Then u **is the minimum of**

$$V(w) = \frac{1}{2}\int_0^L \left(k(x)w'(x)^2 + c(x)w(x)^2\right) dx - \int_0^L f(x)w(x)\,dx - k(L)d_L w(L)$$

over all smooth functions w **satisfying** $w(0) = g_0$**.**

The solution u_h of the variational method minimizes V over $\mathscr{S}_h$. The function V in this case is not interpreted as a potential energy, but rather as a *potential*.

Notice that we have removed the first-order term $b(x)u'(x)$, because if $b(x) \neq 0$ then the bilinear form is no longer symmetric and thus there is no equivalence to a minimization problem.

Example 4.7 Minimization form of 1D fourth-order symmetric problems

Remember the fourth-order problem presented in strong form in Problem 1.5. From Theorem 4.2 we know that **the solution** u **minimizes the function**

$$V(w) = \frac{1}{2}\int_0^L \left(q(x)w''(x)^2 + c(x)w(x)^2\right) dx - \int_0^L f(x)w(x)\,dx - W\,w(L) - T\,w'(L)$$

over all smooth functions w **satisfying** $w(0) = g_0$ **and** $w'(0) = d_0$**.** For simplicity, we have written the boundary conditions at $x = L$ in terms of the force W and torque T (see (1.160).

The solution of the variational method minimizes V over $\mathscr{S}_h$.

When the fourth-order problem models a beam in bending, V is indeed the potential energy of the beam.

Example 4.8 Minimization form of 2D diffusion problems

Let us now consider that u is the solution of Problem 2.1, i.e. that

$$-\text{div}\,(K\nabla u) = f$$

in a 2D domain Ω, with boundary conditions $u = g$ in $\partial\Omega_D$ and $(K\nabla u)\cdot\check{n} = H$ in $\partial\Omega_N$. Then from Theorem 4.2 we conclude that u **minimizes the function**

$$V(w) = \frac{1}{2}\int_\Omega (K\nabla w)\cdot\nabla w\, d\Omega - \int_\Omega f\, w\, d\Omega - \int_{\partial\Omega_N} H\, w\, d\partial\Omega$$

over all functions w **satisfying** $w = g$ **on** $\partial\Omega_D$. Also, that **the solution** u_h **of the variational method minimizes** V **over** $\mathscr{S}_h$.

Bibliography

[1] I. BABUSKA AND A. AZIZ, *On the angle condition in the finite element method*, SIAM J. Numer. Anal, 13 (1976), pp. 214–226.

[2] B. FINLAYSON AND L. SCRIVEN, *The method of weighted residuals—a review*, Appl. Mech. Rev, 19 (1966), pp. 735–748.

[3] B. GALERKIN, *Series in some questions of elastic equilibrium of rods and plates*, Vestnik inzhenerov i tekhnikov, 19 (1915), pp. 897–908.

[4] I. M. GELFAND, R. A. SILVERMAN, ET AL., *Calculus of variations*, Courier Corporation, 2000.

[5] G. H. HARDY, *Weierstrass's non-differentiable function*, Transactions of the American Mathematical Society, 17 (1916), pp. 301–325.

[6] M. KŘÍZEK, *On semi-regular families of triangulations and linear interpolation*, Appl. Math., 36 (1991), pp. 223–232.

[7] W. M. LAI, D. RUBIN, AND E. KREMPL, *Introduction to continuum mechanics*, Butterworth-Heinemann, 2009.

[8] W. RITZ, *Theorie der transversalschwingungen einer quadratischen platte mit freien rändern*, Annalen der Physik, 333 (1909), pp. 737–786.

[9] ——, *Über eine neue methode zur lösung gewisser variationsprobleme der mathematischen physik.*, (1909).

[10] H. SAGAN, *Introduction to the Calculus of Variations*, Courier Corporation, 1992.

[11] S. TIMOSHENKO AND J. GOODIER, *Theory of Elasticity" McGraw-Hill Book Company*, 1951.

Appendix A

Normed Spaces

To be able to assess convergence, or how close our approximations are to the exact solution, we need to define a way to measure distances in a vector space. The most common way to do this is through a *norm*.

Definition A.1 (Norm)**.** *Let V be a vector space. A norm is a function $\|\cdot\|: V \to \mathbb{R}$ such that for $v, u \in V$ and $\alpha \in \mathbb{R}$:*

1. **N.1.** $\|v\| \geq 0$*, and* $\|v\| = 0$ *if and only* $v = 0$.
2. **N.2.** $\|\alpha v\| = |\alpha|\|v\|$.
3. **N.3.** $\|v + u\| \leq \|v\| + \|u\|$ *(triangle inequality).*

The typical norm that you are familiar with is the "Euclidean norm" in $\mathbb{R}^3$. For example, if $x = (x_1, x_2, x_3)$, then $\|x\| = \sqrt{x_1^2 + x_2^2 + x_3^2}$. Clearly if $\|x\| \geq 0$, and if $\|x\| = 0$, then $x = 0$. The second condition, N.2, is also simple to verify, and the triangle inequality is the common statement that the sum of the lengths of two sides of a triangle is always greater or equal than the length of the third. These three conditions are intuitive to understand in the case of $\mathbb{R}^n$, and the fact that the Euclidean norm satisfies them is easy to see. Defining a norm for vector spaces of functions is more delicate, and less intuitive. Let's look at some examples.

Examples:

A.1 For $x = (x_1, x_2, x_3) \in \mathbb{R}^3$, the function $g(x) = \sqrt{2x_1^2 + 3x_2^2 + 4x_3^2}$ is also a norm in $\mathbb{R}^3$. We will not prove this.
The importance of this example is to illustrate that we can endow the same set of vectors with different norms. A simple way to think about this example is that we are using different units to measure distances in each coordinate direction.

A.2 For $v \in V_1 = \{f: [a, b] \to \mathbb{R} \text{ smooth}\}$, we define the *$L^\infty$-norm*

$$\|v\|_{0,\infty} = \max_{x \in [a,b]} |v(x)|. \tag{A.1}$$

Let's check the conditions for this to be a norm, since it is simple in this case. For N.1, since $|v(x)| \geq 0$ for all $x \in [a,b]$, then $\|v\|_{0,\infty} \geq 0$. Also, if $0 = \|v\|_{0,\infty} = \max_{x\in[a,b]} |v(x)| \geq |v(x)| \geq 0$ for any $x \in [a,b]$, then $v = 0$. For N.2,

$$\|\alpha v\|_{0,\infty} = \max_{x\in[a,b]} |\alpha v(x)| = \max_{x\in[a,b]} |\alpha||v(x)| = |\alpha| \max_{x\in[a,b]} |v(x)| = |\alpha|\|v\|_{0,\infty}.$$

Finally, for N.3,

$$\begin{aligned}\|u+v\|_{0,\infty} &= \max_{x\in[a,b]} |u(x)+v(x)| \leq \max_{x\in[a,b]} |u(x)| + |v(x)| \\ &\leq \max_{x\in[a,b]} |u(x)| + \max_{x\in[a,b]} |v(x)| = \|u\|_{0,\infty} + \|v\|_{0,\infty}.\end{aligned}$$

For instance, let $[a,b] = [0,\pi]$, then:

i. If $v(x) = \cos(x)$, then $\|v\|_{0,\infty} = 1$.
ii. If $v(x) = x(x-\pi)$, then $\|v\|_{0,\infty} = -v(\pi/2) = \pi^2/4$.

A.3 For $v \in V_1$ from Example A.2, we define the *L^2-norm*

$$\|v\|_{0,2} = \left[\int_a^b v(x)^2\, dx\right]^{1/2}. \tag{A.2}$$

We will not check that this is a norm, but just state it. For $[a,b] = [0,\pi]$:

i. If $v(x) = \cos(x)$, then $\|v\|_{0,2} = \left[\int_0^\pi \cos(x)^2\, dx\right]^{1/2} = \sqrt{\pi/2}$.
ii. If $v(x) = x(x-\pi)$, then $\|v\|_{0,2} = \left[\int_0^\pi x^2(x-\pi)^2\, dx\right]^{1/2} = \pi^{5/2}/\sqrt{30}$.

A.4 For $v \in V_2 = \{w \in V_1 \mid w(a) = w(b) = 0\}$, we define the *$H^1$-seminorm*

$$|v|_{1,2} = \left[\int_a^b v'(x)^2\, dx\right]^{1/2}. \tag{A.3}$$

The fact that this is a norm requires a discussion of condition N.1: Why does it hold? To answer this, notice that if $|v|_{1,2} = 0$, we can conclude that $v'(x) = 0$ for all $x \in [0,1]$, since the integrand $v'(x)^2 \geq 0$ everywhere. Therefore, $v(x)$ is a constant function. Since $v(a) = 0$, then $v(x) = 0$ for $x \in [a,b]$.

Because of this discussion, $|\cdot|_{1,2}$ is not a norm in the space V_1 in Example A.2, since functions therein need not be zero at the boundaries, and hence condition N.1 is not satisfied. All we would be able to say if $|v|_{1,2} = 0$ is that v is a constant function. For example, let $[a,b] = [0,1]$,

i. If $v(x) = \sin(\pi x)$, $v \in V_2$, then $|v|_{1,2} = \left[\int_0^1 (\pi\cos(\pi x))^2\, dx\right]^{1/2} = \frac{\pi}{\sqrt{2}}$.
ii. If $v(x) = 3$, $v \notin V_2$, then $|v|_{1,2} = \left[\int_0^1 0\, dx\right]^{1/2} = 0$.

A.5 For $v \in V_1$ from Example A.2, we define the H^1*-norm*

$$\begin{aligned} \|v\|_{1,2} &= \left[\int_a^b v(x)^2 \, dx + \int_a^b v'(x)^2 \, dx \right]^{1/2} \\ &= \left[\|v\|_{0,2}^2 + |v|_{1,2}^2 \right]^{1/2}. \end{aligned} \tag{A.4}$$

In contrast to what happens with $|v|_{1,2}$ in Example A.4, condition N.1 is satisfied in this case, since it is satisfied for $\|v\|_{0,2}$.

Notice that we talked about three different norms for space V_1 above: We defined the L^∞-norm, the L^2-norm and the H^1-norm. The three norms measure distance differently, emphasizing different aspects of the functions.

We can now define the notion of a normed space.

Definition A.2 (Normed Space.)**.** *A vector space V with a norm defined over it $\|\cdot\| : V \to \mathbb{R}$ is called a* **normed space***, and denoted by $(V, \|\cdot\|)$.*

Examples:

A.6 The space $\mathbb{R}^n$, $n \in \mathbb{N}$, with the Euclidean norm $\|\cdot\|$ is a normed space $(\mathbb{R}^n, \|\cdot\|)$, since the norm is defined for every element of $\mathbb{R}^n$.

A.7 Consider the space $V_1 = \{f : [a,b] \to \mathbb{R} \text{ smooth}\}$ with the L^∞-norm $\|\cdot\|_{0,\infty}$. Since all smooth functions are bounded in $[a,b]$, the L^∞-norm is well defined for every function in V_1, and hence $(V_1, \|\cdot\|_{0,\infty})$ is a normed space.

A.8 Consider again the space $V_1 = \{f : [a,b] \to \mathbb{R} \text{ smooth}\}$ with the L^2-norm $\|\cdot\|_{0,2}$. Since all smooth functions are bounded in $[a,b]$, the integrals needed to compute the L^2-norm are well defined for every function in V_1, and hence $(V_1, \|\cdot\|_{0,2})$ is a normed space. It is, however, a different normed space than $(V_1, \|\cdot\|_{0,\infty})$, since functions that are close in one may not be close in the other, as we shall see.

A.9 Consider the space $V_2 = \{f : (a,b) \to \mathbb{R} \text{ smooth}\}$ (notice the open interval) with the L^2-norm $\|\cdot\|_{0,2}$. The function $f(x) = 1/(x-a)$ is in V_2, since it is smooth in (a,b), but

$$\|f\|_{0,2}^2 = \int_a^b \frac{1}{(x-a)^2} \, dx = +\infty,$$

so the norm is not defined for f. Therefore, $(V_2, \|\cdot\|_{0,2})$ is *not* a normed space.

A.10 Let $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$. For such domain Ω, the norm $\|v\|_{0,2}$ of $v : \Omega \to \mathbb{R}$ is defined as

$$\|v\|_{0,2} = \left[\int_\Omega v(x)^2 \, d\Omega \right]^{1/2}. \tag{A.5}$$

The set

$$L^2(\Omega) = \{v\colon \Omega \to \mathbb{R} \mid \|v\|_{0,2} < \infty\} \tag{A.6}$$

is called the $L^2(\Omega)$ **space**, and $(L^2(\Omega), \|\cdot\|_{0,2})$ is a normed space. The space $L^2(\Omega)$ is said to contain all *square-integrable* functions, and these need not be smooth. For example, if $\Omega = [-1,1]$, it contains the Heaviside function

$$H(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0. \end{cases}$$

In contrast, $H(x) \notin L^2(\mathbb{R})$, since $\|H\|_{0,2} = \infty$ in this case.

A.11 Let $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$. For such domain Ω, we define the H^1*-norm* as

$$\|v\|_{1,2} = \left[\|v\|_{0,2}^2 + \sum_{i=1}^{n} \left\| \frac{\partial v}{\partial x_i} \right\|_{0,2}^2 \right]^{1/2}.$$

With it, we can define the $H^1(\Omega)$**-space** as

$$H^1(\Omega) = \left\{v\colon \Omega \to \mathbb{R} \mid \|v\|_{1,2} < \infty\right\}, \tag{A.7}$$

and $(H^1(\Omega), \|\cdot\|_{1,2})$ is normed space. Functions in $H^1(\Omega)$ contain all functions in which both the function and each one of its partial derivatives is square integrable. Alternatively, the function and each one of its partial derivatives is in $L^2(\Omega)$. Therefore, if a funtion $v \in H^1(\Omega)$, then $v \in L^2(\Omega)$. For example: Let $\Omega = [-1,1] \times [-1,1]$, then

☞ The definition of H^1 here is incomplete. We will have an opportunity to complete the definition later.

i. The function $v(x_1, x_2) = x_1^2 + x_2^3 \in H^1(\Omega)$, since

$$\|v\|_{1,2}^2 = \int_{-1}^{1}\int_{-1}^{1} (x_1^2 + x_2^3)^2 \, dx_1 dx_2 + \int_{-1}^{1}\int_{-1}^{1} (2x_1)^2 \, dx_1 dx_2 + \int_{-1}^{1}\int_{-1}^{1} (3x_2^2)^2 \, dx_1 dx_2 = \frac{292}{21} < \infty.$$

ii. The function $v(x_1, x_2) = \ln(1+x_1) + \ln(1+x_2) \notin H^1(\Omega)$, but $v \in L^2(\Omega)$, since

$$\|v\|_{0,2}^2 = \int_{-1}^{1}\int_{-1}^{1} (\ln(1+x_1) + \ln(1+x_2))^2 \, dx_1 dx_2 = 24 + 8\ln(4)(\ln(2) - 2) < \infty.$$

$$\|v\|_{1,2}^2 = \|v\|_{0,2}^2 + \int_{-1}^{1} \frac{1}{(1+x_1)^2} \, dx_1 dx_2 + \int_{-1}^{1} \frac{1}{(1+x_2)^2} \, dx_1 dx_2 = \infty.$$

A.12 Let $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$. For $v \in H^1(\Omega)$ we can define the H^1*-seminorm* as

$$|v|_{1,2} = \left[\int_\Omega \|\nabla v\|^2 \, d\Omega \right]^{1/2}. \tag{A.8}$$

Notice that

$$|v|_{1,2}^2 = \int_\Omega \|\nabla v\|^2 \, d\Omega = \int_\Omega \sum_{i=1}^n \left| \frac{\partial v}{\partial x_i} \right|^2 \, d\Omega = \sum_{i=1}^n \int_\Omega \left| \frac{\partial v}{\partial x_i} \right|^2 \, d\Omega$$
$$= \sum_{i=1}^n \left\| \frac{\partial v}{\partial x_i} \right\|_{0,2}^2 .$$

This allows us to write the H^1-norm as

$$\|v\|_{1,2}^2 = \|v\|_{0,2}^2 + |v|_{1,2}^2. \tag{A.9}$$

A.13 Let $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$ and $m \in \mathbb{N}_0 = \{0\} \cup \mathbb{N}$. The H^m-seminorm of a function $u\colon \Omega \to \mathbb{R}$ is defined as

$$|u|_{m,2}^2 = \sum_{\substack{i_1,\dots,i_n \geq 0 \\ i_1+\dots+i_n=m}} \left\| \frac{\partial^m u}{\partial x_1^{i_1} \dots \partial x_n^{i_n}} \right\|_{0,2}^2 , \tag{A.10}$$

where $\partial^m u / \partial x_k^0 = u$.
For example, the H^2-seminorm in $\mathbb{R}^2$ is

$$|u|_{2,2}^2 = \left\| \frac{\partial^2 u}{\partial x_1^2} \right\|_{0,2}^2 + 2 \left\| \frac{\partial^2 u}{\partial x_1 \partial x_2} \right\|_{0,2}^2 + \left\| \frac{\partial^2 u}{\partial x_2^2} \right\|_{0,2}^2 ,$$

and the H^0-seminorm is directly the L^2-norm.
The H^m-norm is then defined as

$$\|u\|_{m,2}^2 = \sum_{i=0}^m |u|_{i,2}^2 \tag{A.11}$$

With it, we can define the $H^m(\Omega)$-**space** as

$$H^m(\Omega) = \left\{ v\colon \Omega \to \mathbb{R} \mid \|u\|_{m,2} < \infty \right\}. \tag{A.12}$$