Chapter 10
Finite Elements and Weak Solutions

In Chapter 5, we studied the oldest, and in many ways the simplest, class of numerical algorithms for approximating the solutions to partial differential equations: those based on finite difference approximations. In the present chapter, we introduce the second of the two major numerical paradigms: the finite element method. Finite elements are of more recent vintage, having first appeared soon after the Second World War; historical details can be found in [**113**]. As a consequence of their ability to adapt to complicated geometries, finite elements have, in many situations, become the method of choice for solving equilibrium boundary value problems governed by elliptic partial differential equations. Finite elements can also be adapted to dynamical problems, but lack of space prevents us from pursuing such extensions in this text.

Finite elements rely on a more sophisticated understanding of the partial differential equation, in that, unlike finite differences, they are not obtained by simply replacing derivatives by their numerical approximations. Rather, they are initially founded on an associated minimization principle that, as we learned in Chapter 9, characterizes the unique solution to a positive definite boundary value problem. The basic idea is to restrict the minimizing functional to an appropriately chosen finite-dimensional subspace of functions. Such a restriction produces a finite-dimensional minimization problem, which can then be solved by numerical linear algebra. When properly formulated, the restricted finite-dimensional minimization problem will have a solution that well approximates the true minimizer, and hence the solution to the original boundary value problem. To gain familiarity with the underlying principles, we will first illustrate the basic constructions in the context of boundary value problems for ordinary differential equations. The following section extends finite element analysis to boundary value problems associated with the two-dimensional Laplace and Poisson equations, thereby revealing the key features used in applications to the numerical solution of multidimensional equilibrium boundary value problems.

An alternative approach to the finite element method, one that can be applied even in situations in which no minimum principle is available, is founded on the concept of a weak solution to the differential equation, a construction of independent analytical importance. The term "weak" refers to the fact that one is able to relax the differentiability requirements imposed on classical solutions. Indeed, as we will show, discontinuous shock wave solutions as well as the nonsmooth, and hence nonclassical, solutions to the wave equation that we encountered in Chapters 2 and 4 can all be rigorously characterized through the weak solution formulation. For the finite element approximation, rather than impose the weak solution criterion on the entire infinite-dimensional function space, one again restricts to a

P.J. Olver, *Introduction to Partial Differential Equations*, Undergraduate Texts in Mathematics,
DOI 10.1007/978-3-319-02099-0_10,

suitably chosen finite-dimensional subspace. For positive definite boundary value problems, which necessarily admit a minimization principle, the weak solution approach leads to the same finite element equations.

A rigorous justification and proof of convergence of the finite element approximations requires further analysis, and we refer the interested reader to more specialized texts, such as [**6**, **113**, **126**]. In this chapter, we shall focus our effort on understanding how to formulate and implement the finite element method in practical contexts.

10.1 Minimization and Finite Elements

To explain the principal ideas underpinning the finite element method, we return to the abstract framework for boundary value problems that was developed in Chapter 9. Recall Theorem 9.26, which characterizes the unique solution to a positive definite linear system as the minimizer, $u_\star \in U$, of an associated quadratic functional $Q\colon U \to \mathbb{R}$. For boundary value problems governed by differential equations, U is an infinite-dimensional function space containing all sufficiently smooth functions that satisfy the prescribed homogeneous boundary conditions. (Modifications to deal with inhomogeneous boundary conditions will be discussed in due course.)

This framework sets the stage for the first key idea of the finite element method. Instead of trying to minimize the functional $Q[u]$ over the entire infinite-dimensional function space, we will seek to minimize it over a *finite-dimensional subspace* $W \subset U$. The effect is to reduce a problem in analysis — a boundary value problem for a differential equation — to a problem in linear algebra, and hence one that a computer is capable of solving. On the surface, the idea seems crazy: how could one expect to come close to finding the minimizer in a gigantic infinite-dimensional function space by restricting the search to a mere finite-dimensional subspace? But this is where the magic of infinite dimensions comes into play. One can, in fact, approximate all (reasonable) functions arbitrarily closely by functions belonging to finite-dimensional subspaces. Indeed, you are already familiar with two examples: Fourier series, where one approximates rather general periodic functions by trigonometric polynomials, and interpolation theory, in which one approximates functions by ordinary polynomials, or, more sophisticatedly, by splines, [**89**, **102**]. Thus, the finite element idea perhaps is not as outlandish as it might initially seem.

To be a bit more explicit, let us begin with a linear operator $L\colon U \to V$ between real inner product spaces, where, as in Section 9.1, $\langle\, u\,, \widetilde{u}\,\rangle$ is used to denote the inner product in U, and $\langle\!\langle\, v\,, \widetilde{v}\,\rangle\!\rangle$ the inner product in V. To ensure uniqueness of solutions, we always assume that L has trivial kernel: $\ker L = \{0\}$. According to Theorem 9.26, the element $u_\star \in U$ that minimizes the quadratic function(al)

$$Q[u] = \tfrac{1}{2}\, |||\, L[u]\, |||^2 - \langle\, f\,, u\,\rangle, \tag{10.1}$$

where $|||\cdot|||$ denotes the norm in V, is the solution to the linear system

$$S[u] = f, \qquad \text{where} \qquad S = L^* \circ L, \tag{10.2}$$

with $L^*\colon V \to U$ denoting the adjoint operator. The hypothesis that L has trivial kernel implies that S is a self-adjoint positive definite linear operator, which implies that the solution to (10.2), and hence the minimizer of $Q[u]$, is unique. In our applications, L is a linear differential operator between function spaces, e.g., the gradient, while $Q[u]$ represents a quadratic functional, e.g., the Dirichlet principle, and the associated linear

system (10.2) forms a positive definite boundary value problem, e.g., the Poisson equation along with suitable boundary conditions.

To form a finite element approximation to the solution $u_\star \in U$, rather than try to minimize $Q[u]$ on the entire function space U, we now seek to minimize it on a suitably chosen finite-dimensional subspace $W \subset U$. We will specify W by selecting a set of linearly independent functions $\varphi_1, \dots, \varphi_n \in U$, and letting W be their span. Thus, $\varphi_1, \dots, \varphi_n$ form a basis of W, whereby $\dim W = n$, and the general element of W is a (uniquely determined) linear combination

$$w(x) = c_1\varphi_1(x) + \cdots + c_n\varphi_n(x) \tag{10.3}$$

of the basis functions. Our goal is to minimize $Q[w]$ over all possible $w \in W$; in other words, we need to determine the coefficients $c_1, \dots, c_n \in \mathbb{R}$ such that

$$Q[w] = Q[c_1\varphi_1 + \cdots + c_n\varphi_n] \tag{10.4}$$

is as small as possible. Substituting (10.3) back into (10.1) and then expanding, using the linearity of L and then the bilinearity of the inner product, we find that the resulting expression is the quadratic function

$$P(\mathbf{c}) = \frac{1}{2}\sum_{i,j=1}^{n} k_{ij}\, c_i\, c_j - \sum_{i=1}^{n} b_i\, c_i = \tfrac{1}{2}\,\mathbf{c}^T K\,\mathbf{c} - \mathbf{c}^T\mathbf{b}, \tag{10.5}$$

in which

- $\mathbf{c} = (\, c_1, c_2, \dots, c_n \,)^T \in \mathbb{R}^n$ is the vector of unknown coefficients in (10.3);
- $K = (k_{ij})$ is the symmetric $n \times n$ matrix with entries

$$k_{ij} = \langle\!\langle\, L[\varphi_i]\,, L[\varphi_j]\,\rangle\!\rangle, \qquad i, j = 1, \dots, n; \tag{10.6}$$

- $\mathbf{b} = (\, b_1, b_2, \dots, b_n \,)^T$ is the vector with entries

$$b_i = \langle\, f\,, \varphi_i\,\rangle, \qquad i = 1, \dots, n. \tag{10.7}$$

Note that formula (10.6) uses the inner product on the target space V, whereas (10.7) relies on the inner product on the domain space U.

Thus, once we specify the basis functions φ_i, the coefficients k_{ij} and b_i are all known quantities. We have effectively reduced our original problem to the finite-dimensional problem of minimizing the quadratic function (10.5) over all possible vectors $\mathbf{c} \in \mathbb{R}^n$. The symmetric matrix K is, in fact, positive definite, since, by the preceding computation,

$$\mathbf{c}^T K\,\mathbf{c} = \sum_{i,j=1}^{n} k_{ij}\, c_i\, c_j = |||\, L[c_1\,\varphi_1(x) + \cdots + c_n\,\varphi_n]\,|||^2 = |||\, L[w]\,|||^2 > 0, \tag{10.8}$$

as long as $L[w] \neq 0$. Moreover, our initial assumption tells us that $L[w] = 0$ if and only if $w = 0$, which, by linear independence, occurs only when $\mathbf{c} = \mathbf{0}$. Thus, (10.8) is indeed positive for all $\mathbf{c} \neq \mathbf{0}$. We can now invoke the finite-dimensional minimization result contained in Example 9.25 to conclude that the unique minimizer to (10.5) is obtained by solving the associated linear system

$$K\,\mathbf{c} = \mathbf{b}, \qquad \text{whereby} \qquad \mathbf{c} = K^{-1}\mathbf{b}. \tag{10.9}$$

Remark: When of moderate size, the linear system (10.9) can be solved by basic Gaussian Elimination. When the size (i.e., the dimension, n, of the subspace W) becomes too large, as is often the case in dealing with partial differential equations, it is better to rely on an iterative linear system solver, e.g., Gauss–Seidel or Successive Over–Relaxation (SOR); see [**89**, **118**] for details.

This summarizes the basic abstract setting for the finite element method. The key issue, then, is how to effectively choose the finite-dimensional subspace W. Two candidates that might spring to mind are the space of polynomials of degree $\leq n$ and the space of trigonometric polynomials (truncated Fourier series) of degree $\leq n$. However, for a variety of reasons, neither is well suited to the finite element method. One constraint is that the functions in W must satisfy the relevant boundary conditions — otherwise, W would not be a subspace of U. More importantly, in order to obtain sufficient accuracy of the approximate solution, the linear algebraic system (10.9) will typically — especially when dealing with partial differential equations — be quite large, and hence it is desirable that the coefficient matrix K be as sparse as possible, i.e., have lots of zero entries. Otherwise, computing the solution may well be too time-consuming to be of much practical value.

With this in mind, the second innovative contribution of the finite element method is to first (paradoxically) *enlarge* the space U of allowable functions upon which to minimize the quadratic functional $Q[u]$. The governing differential equation requires its (classical) solutions to have a certain degree of smoothness, whereas the associated minimization principle typically requires that they possess only half as many derivatives. Thus, for second-order boundary value problems, the differential equation requires continuous second-order derivatives, while the quadratic functional $Q[u]$ involves only first-order derivatives. It fact, it can be rigorously shown that, under rather mild hypotheses, the functional retains the *same* minimizing solution, even when one allows functions that fail to qualify as classical solutions to the differential equation. We will proceed to develop the method in the context of particular, fairly elementary examples.

Exercises

10.1.1. Let $U = \{\, u(x) \in \mathrm{C}^2[0,\pi] \,|\, u(0) = u(\pi) = 0 \,\}$ and $V = \{\, v(x) \in \mathrm{C}^1[0,\pi] \,\}$ both be equipped with the L^2 inner product. Let $L\colon U \to V$ be given by $L[u] = D[u] = u'$, and $f(x) = x - 1$. (*a*) Write out the quadratic functional $Q[u]$ given by (10.1). (*b*) Write out the associated boundary value problem (10.2). (*c*) Find the function $u_\star(x) \in U$ that minimizes $Q[u]$. What is the value of $Q[u_\star]$? (*d*) Let $W \subset U$ be the subspace spanned by $\sin x$ and $\sin 2x$. Write out the corresponding finite-dimensional minimization problem (10.8). (*e*) Find the function $w_\star(x) \in W$ that minimizes $Q[w]$. Is $Q[w_\star] \geq Q[u_\star]$? If not, why not? How close is your finite element minimizer $w_\star(x)$ to the actual minimizer $u_\star(x)$?

10.1.2. Let $U = \{\, u(x) \in \mathrm{C}^2[0,1] \,|\, u(0) = u(1) = 0 \,\}$ and $V = \{\, v(x) \in \mathrm{C}^1[0,1] \,\}$ both have the L^2 inner product. Let $L\colon U \to V$ be given by $L[u] = u'(x) - u(x)$, and $f(x) = 1$ for all x. (*a*) Write out the quadratic functional $Q[u]$ given by (10.1). (*b*) Write out the associated boundary value problem (10.2). (*c*) Find the function $u_\star(x) \in U$ that minimizes $Q[u]$. What is the value of $Q[u_\star]$? (*d*) Let $W \subset U$ be the subspace containing all cubic polynomials $p(x)$ that satisfy the boundary conditions: $p(0) = p(1) = 0$. Find a basis of W and then write out the corresponding finite-dimensional minimization problem (10.8). (*e*) Find the polynomial $p_\star(x) \in W$ that minimizes $Q[p]$ for $p \in W$. Is $Q[p_\star] \geq Q[u_\star]$? If not, why not? How close is your finite element minimizer $p_\star(x)$ to the minimizer $u_\star(x)$?

10.1.3. Let $U = \{\, u(x) \in \mathrm{C}^2[1,2] \,|\, u(1) = u(2) = 0 \,\}$, $V = \{\, (v_1(x), v_2(x))^T \,|\, v_1, v_2 \in \mathrm{C}^1[1,2] \,\}$, both be endowed with the L^2 inner product. Let $L\colon U \to V$ be given by $L[u] = \begin{pmatrix} x\,u'(x) \\ \sqrt{2}\,u(x) \end{pmatrix}$, and let $f(x) = 2$ for all $1 \le x \le 2$. (*a*) Write out the quadratic functional $Q[u]$ given by (10.1). (*b*) Write out the associated boundary value problem (10.2). (*c*) Find the function $u_\star(x) \in U$ that minimizes $Q[u]$. What is the value of $Q[u_\star]$? (*d*) Let $W \subset U$ be the subspace containing all cubic polynomials $p(x)$ that satisfy the boundary conditions $p(1) = p(2) = 0$. Find a basis of W and then write out the corresponding finite-dimensional minimization problem (10.8). (*e*) Find the polynomial $p_\star(x) \in W$ that minimizes $Q[p]$ for $p \in W$. Is $Q[p_\star] \ge Q[u_\star]$? If not, why not? How close is your finite element minimizer $p_\star(x)$ to the actual minimizer $u_\star(x)$?

♡ 10.1.4. (*a*) Find the solution to the boundary value problem $-u'' = x^2 - x$, $u(-1) = u(1) = 0$. (*b*) Write down a quadratic functional $Q[u]$ that is minimized by your solution. (*c*) Let W be the subspace spanned by the two functions $(1-x^2)$, $x\,(1-x^2)$. Find the function $w_\star(x) \in W$ that minimizes the restriction of your quadratic functional to W. Compare $w_\star$ with your solution from part (*a*). (*d*) Answer part (*c*) for the subspace W spanned by $\sin \pi x, \sin 2\pi x$. Which of the two approximations is the better?

♡ 10.1.5. (*a*) Find the function $u_\star(x)$ that minimizes $Q[u] = \displaystyle\int_0^1 \Big[\, \tfrac{1}{2}\,(x+1)\,u'(x)^2 - u(x) \,\Big]\, dx$ over the vector space U consisting of C^2 functions satisfying $u(0) = u(1) = 0$. (*b*) Let $W_3 \subset U$ be the subspace consisting of all cubic polynomials $w(x)$ that satisfy the same boundary conditions. Find the function $w_\star(x)$ that minimizes the restriction $Q[w]$ for $w \in W_3$. Compare $w_\star(x)$ and $u_\star(x)$: how close are they in the L^2 norm? What is the maximal discrepancy $|\, w_\star(x) - u_\star(x) \,|$ for $0 \le x \le 1$? (*c*) Suppose you enlarge your finite-dimensional subspace $W_4 \subset U$ to contain all quartic polynomials that satisfy the boundary conditions. Is your new finite element approximation better? Discuss.

♡ 10.1.6. (*a*) Find the function $u_\star(x)$ that minimizes $Q[u] = \displaystyle\int_0^1 \Big[\, \tfrac{1}{2}\,e^x\,u'(x)^2 - 3\,u(x) \,\Big]\, dx$ over the space U consisting of C^2 functions satisfying the boundary conditions $u(0) = u'(1) = 0$. (*b*) Let $W \subset U$ be the subspace containing all cubic polynomials $w(x)$ that satisfy the boundary conditions. Find the polynomial $w_\star(x)$ that minimizes the restriction $Q[w]$ for $w \in W$. Compare $w_\star(x)$ and $u_\star(x)$: how close are they in the L^2 norm? What is the maximal discrepancy $|\, w_\star(x) - u_\star(x) \,|$ for $0 \le x \le 1$?

10.1.7. Consider the Dirichlet boundary value problem

$$-\Delta u = x\,(1-x) + y\,(1-y), \qquad u(x,0) = u(x,1) = u(0,y) = u(1,y) = 0,$$

on the unit square $\{\, 0 < x, y < 1 \,\}$.

(*a*) Find the exact solution $u_\star(x,y)$. *Hint*: It is a polynomial.

(*b*) Write down a minimization principle $Q[u]$ that characterizes the solution. Be careful to specify the function space U over which the minimization takes place.

(*c*) Let $W \subset U$ be the subspace spanned by the four functions $\sin \pi x \sin \pi y$, $\sin 2\pi x \sin \pi y$, $\sin \pi x \sin 2\pi y$, and $\sin 2\pi x \sin 2\pi y$. Find the function $w_\star \in W$ that minimizes the restriction of $Q[w]$ to $w \in W$. How close is $w_\star$ to the solution you found in part (*a*)?

♢ 10.1.8. Justify the identification of (10.4) with the quadratic function (10.5).

10.2 Finite Elements for Ordinary Differential Equations

To understand the preceding abstract formulation in concrete terms, let us focus our attention on boundary value problems governed by a second-order ordinary differential equation.

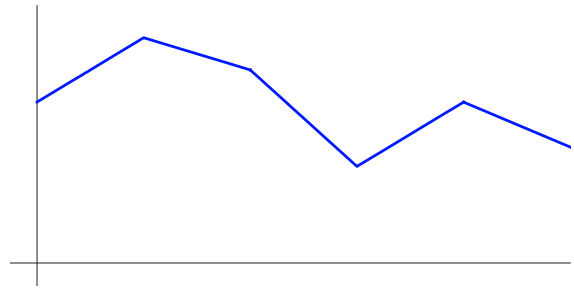


Figure 10.1. A continuous piecewise affine function.

For example, we might be interested in solving a Sturm–Liouville problem (9.71) subject to, say, homogeneous Dirichlet boundary conditions. Once we understand how the finite element constructions work in this relatively simple context, we will be in a good position to extend the techniques to much more general linear boundary value problems governed by elliptic partial differential equations.

For such one-dimensional boundary value problems, a popular and effective choice of the finite-dimensional subspace W is to employ continuous, piecewise affine functions. Recall that a function is *affine* if its graph is a straight line: $f(x) = a\,x + b$. (The function is *linear*, in accordance with Definition B.32, if and only if $b = 0$.) A function is called *piecewise affine* if its graph consists of a finite number of straight line segments; a typical example is plotted in Figure 10.1. Continuity requires that the individual segments be connected together end to end.

Given a boundary value problem on a bounded interval $[a, b]$, let us fix a finite collection of *nodes*

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

The formulas simplify if one uses equally spaced nodes, but this is not necessary for the construction to be carried out. Let W denote the vector space consisting of all continuous functions $w(x)$ that are defined on the interval $a \le x \le b$, satisfy the homogeneous boundary conditions, and are affine when restricted to each subinterval $[x_j, x_{j+1}]$. On each subinterval, we write

$$w(x) = c_j + b_j(x - x_j), \qquad \text{for} \qquad x_j \le x \le x_{j+1}, \qquad j = 0, \dots, n-1,$$

for certain constants c_j, b_j. Continuity of $w(x)$ requires

$$c_j = w(x_j^+) = w(x_j^-) = c_{j-1} + b_{j-1}\,h_{j-1}, \qquad j = 1, \dots, n-1, \tag{10.10}$$

where $h_{j-1} = x_j - x_{j-1}$ denotes the length of the j^{th} subinterval. The homogeneous Dirichlet boundary conditions at the endpoints require

$$w(a) = c_0 = 0, \qquad w(b) = c_{n-1} + b_{n-1}\,h_{n-1} = 0. \tag{10.11}$$

Observe that the function $w(x)$ involves a total of $2\,n$ unspecified coefficients $c_0, \dots, c_{n-1}$, $b_0, \dots, b_{n-1}$. The continuity conditions (10.10) and the second boundary condition (10.11) uniquely determine the b_j. The first boundary condition specifies c_0, while the remaining $n-1$ coefficients $c_1 = w(x_1), \dots, c_{n-1} = w(x_{n-1})$ are arbitrary, specifying the values of $w(x)$ at the interior nodes. We conclude that the finite element subspace W has dimension $n-1$, the number of interior nodes.

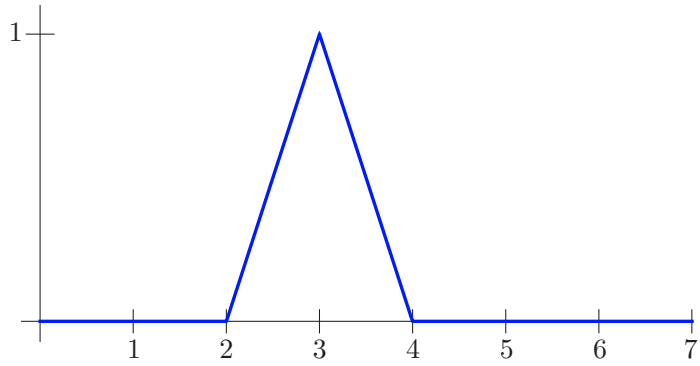


Figure 10.2. A hat function.

Remark: Every function $w(x)$ in our subspace has piecewise constant first derivative $w'(x)$. However, the jump discontinuities in $w'(x)$ imply that its second derivative $w''(x)$ may well include delta function impulses at the nodes, and hence $w(x)$ is far from being a solution to the differential equation. Nevertheless, in practice, the finite element minimizer $w_\star(x) \in W$ will (under suitable assumptions) provide a reasonable approximation to the actual solution $u_\star(x)$.

The most convenient basis for W consists of the *hat functions*, which are continuous, piecewise affine functions satisfying

$$\varphi_j(x_k) = \begin{cases} 1, & j = k, \\ 0, & j \neq k, \end{cases} \qquad \text{for} \qquad j = 1, \dots, n-1, \qquad k = 0, \dots, n. \tag{10.12}$$

The graph of a typical hat function appears in Figure 10.2. The explicit formula is easily established:

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}}, & x_{j-1} \leq x \leq x_j, \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j}, & x_j \leq x \leq x_{j+1}, \\ 0, & x \leq x_{j-1} \ \text{ or } \ x \geq x_{j+1}, \end{cases} \qquad j = 1, \dots, n-1. \tag{10.13}$$

One advantage of using these basis functions is that, thanks to (10.12), the coefficients in the linear combination

$$w(x) = c_1 \varphi_1(x) + \cdots + c_n \varphi_n(x)$$

coincide with its values at the nodes:

$$c_j = w(x_j), \qquad j = 1, \dots, n. \tag{10.14}$$

Example 10.1. Let $\kappa(x) > 0$ for $0 \leq x \leq \ell$. Consider the equilibrium equations

$$S[u] = -\frac{d}{dx}\left(\kappa(x)\,\frac{du}{dx}\right) = f(x), \qquad 0 < x < \ell, \qquad u(0) = u(\ell) = 0,$$

for a nonuniform bar with fixed ends and variable stiffness $\kappa(x)$, that is subject to an external forcing $f(x)$. In order to find a finite element approximation to the resulting

displacement $u(x)$, we begin with the minimization principle based on the quadratic functional

$$Q[u] = \int_0^\ell \left[\tfrac{1}{2}\kappa(x)\,u'(x)^2 - f(x)\,u(x) \right] dx,$$

which is a special case of (9.75). We divide the interval $[0, \ell]$ into n equal subintervals, each of length $h = \ell/n$. The resulting uniform mesh has nodes

$$x_j = j\,h = \frac{j\,\ell}{n}, \qquad\qquad j = 0, \dots, n.$$

The corresponding finite element basis hat functions are explicitly given by

$$\varphi_j(x) = \begin{cases} (x - x_{j-1})/h, & x_{j-1} \le x \le x_j, \\ (x_{j+1} - x)/h, & x_j \le x \le x_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \qquad\qquad j = 1, \dots, n-1. \tag{10.15}$$

The associated linear system (10.9) has coefficient matrix entries

$$k_{ij} = \langle\!\langle\, \varphi_i' \,,\, \varphi_j' \,\rangle\!\rangle = \int_0^\ell \varphi_i'(x)\,\varphi_j'(x)\,\kappa(x)\,dx, \qquad\qquad i, j = 1, \dots, n-1.$$

Since the function $\varphi_i(x)$ vanishes except on the interval $x_{i-1} < x < x_{i+1}$, while $\varphi_j(x)$ vanishes outside $x_{j-1} < x < x_{j+1}$, the integral will vanish unless $i = j$ or $i = j \pm 1$. Moreover,

$$\varphi_i'(x) = \begin{cases} 1/h, & x_{j-1} \le x \le x_j, \\ -1/h, & x_j \le x \le x_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \qquad\qquad j = 1, \dots, n-1.$$

Therefore, the finite element coefficient matrix assumes the tridiagonal form

$$K = \frac{1}{h^2} \begin{pmatrix} s_0 + s_1 & -s_1 & & & & \\ -s_1 & s_1 + s_2 & -s_2 & & & \\ & -s_2 & s_2 + s_3 & -s_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -s_{n-3} & s_{n-3} + s_{n-2} & -s_{n-2} \\ & & & & -s_{n-2} & s_{n-2} + s_{n-1} \end{pmatrix}, \tag{10.16}$$

where

$$s_j = \int_{x_j}^{x_{j+1}} \kappa(x)\,dx \tag{10.17}$$

is the total stiffness of the j^{th} subinterval. The corresponding right-hand side has entries

$$\begin{aligned} b_j = \langle\, f \,,\, \varphi_j \,\rangle &= \int_0^\ell f(x)\,\varphi_j(x)\,dx \\ &= \frac{1}{h}\left[\int_{x_{j-1}}^{x_j} (x - x_{j-1}) f(x)\,dx + \int_{x_j}^{x_{j+1}} (x_{j+1} - x) f(x)\,dx \right]. \end{aligned} \tag{10.18}$$

In practice, we do not have to explicitly evaluate the integrals (10.17, 18), but may replace them by suitably close numerical approximations. When the step size $h \ll 1$ is small, then

the integrals are taken over small intervals, and so the elementary trapezoid rule, [**24**, **108**], produces sufficiently accurate approximations:

$$s_j \;\approx\; \frac{h}{2}\,[\,\kappa(x_j) + \kappa(x_{j+1})\,], \qquad\qquad b_j \;\approx\; h\,f(x_j). \tag{10.19}$$

The resulting finite element system $K\,\mathbf{c} = \mathbf{b}$ is then solved for $\mathbf{c}$, whose entries, according to (10.14), coincide with the values of the finite element approximation to the solution at the nodes: $c_j = w(x_j) \approx u(x_j)$. Indeed, the tridiagonal Gaussian Elimination algorithm, [**89**], will rapidly produce the desired solution. Since the accuracy of the finite element solution increases with the number of nodes, this numerical scheme allows us to easily compute very accurate approximations to the solution to the boundary value problem.

In particular, in the homogeneous case $\kappa(x) \equiv 1$, the coefficient matrix (10.16) reduces to the special form

$$K = \frac{1}{h}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}. \tag{10.20}$$

In this case, the j^{th} equation in the finite element linear system is, upon dividing by h,

$$-\,\frac{c_{j+1} - 2\,c_j + c_{j-1}}{h^2} \;=\; f(x_j). \tag{10.21}$$

Since $c_j \approx u(x_j)$, the left-hand side coincides with the standard finite difference approximation to minus the second derivative $-u''(x_j)$ at the node x_j, cf. (5.5). As a result, in this particular case the finite element and finite difference numerical solution schemes happen to coincide.

The sparse tridiagonal nature of the finite element matrix is a consequence of the fact that the basis functions are zero on much of the interval, or, in more mathematical language, that they have *small support*, in the following sense.

Definition 10.2. The *support* of a function $f(x)$, written $\operatorname{supp} f$, is the closure of the set where $f(x) \neq 0$.

Thus, a point x will belong to the support, provided f is not zero there, or at least is not zero at nearby points. For example, the support of the hat function (10.13) is the (small) interval $[x_{j-1}, x_{j+1}]$. The key property, ensuring sparseness, is that the integral of the product of two functions will be zero if their supports have empty intersection, or, slightly more generally, have only a finite number of points in common.

Example 10.3. Consider the boundary value problem

$$-\,\frac{d}{dx}\,(x+1)\,\frac{du}{dx} = 1, \qquad\qquad u(0) = 0, \qquad\qquad u(1) = 0. \tag{10.22}$$

The explicit solution is easily found by direct integration:

$$u(x) = -\,x + \frac{\log(x+1)}{\log 2}\,. \tag{10.23}$$

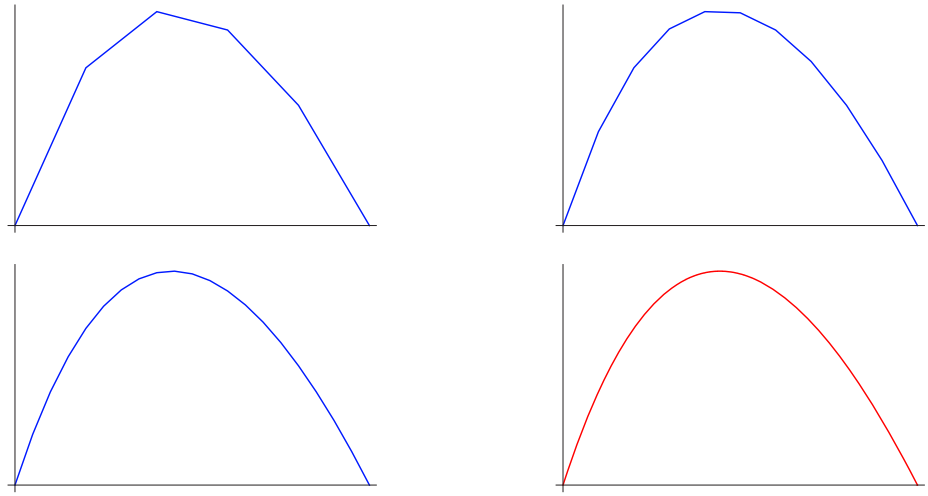


Figure 10.3. Finite element solution to (10.22).

It minimizes the associated quadratic functional

$$Q[u] = \int_0^\ell \left[\tfrac{1}{2}(x+1)\,u'(x)^2 - u(x) \right] dx \tag{10.24}$$

over the space of all C^2 functions $u(x)$ that satisfy the given boundary conditions. The finite element system (10.9) has coefficient matrix given by (10.16) and right-hand side (10.18), where

$$s_j = \int_{x_j}^{x_{j+1}} (1+x)\,dx = h\,(1+x_j) + \tfrac{1}{2}h^2 = h + h^2\big(j+\tfrac{1}{2}\big), \qquad b_j = \int_{x_j}^{x_{j+1}} 1\,dx = h.$$

The resulting piecewise affine approximation to the solution is plotted in Figure 10.3. The first three graphs contain, respectively, 5, 10, 20 nodes, so that $h = .2, .1, .05$, while the last plots the exact solution (10.23). The maximal errors at the nodes are, respectively, $.000298, .000075, .000019$, while the maximal overall errors between the exact solution and its piecewise affine finite element approximations are $.00611, .00166, .00043$. (One can more closely fit the solution curve by employing a cubic spline to interpolate the computed nodal values, [**89**, **102**], which has the effect of reducing the preceding maximal overall errors by a factor of, approximately, 20.) Thus, even when computed on rather coarse meshes, the finite element approximation gives quite respectable results.

Remark: One can obtain a smoother, and hence more realistic, approximation to the solution by smoothly interpolating the finite element approximations $c_j \approx u(x_j)$ at the nodes, e.g., by use of cubic splines, [**89**, **102**]. Alternatively, one can require that the finite element functions themselves be smoother, e.g., by making the finite element subspace consist of piecewise cubic splines that satisfy the boundary conditions.

Exercises

♣ 10.2.1. Use the finite element method to approximate the solution to the boundary value problem $-\dfrac{d}{dx}\left(e^{-x}\dfrac{du}{dx}\right) = 1,\ u(0) = u(2) = 0$. Carefully explain how you are setting up the calculation. Plot the resulting solutions and compare your answer with the exact solution. You should use an equally spaced mesh, but try at least three different mesh spacings and compare your results. By inspecting the errors in your various approximations, can you predict how many nodes would be required for six-digit accuracy of the numerical approximation?

♠ 10.2.2. For each of the following boundary value problems: (*i*) Solve the problem exactly. (*ii*) Approximate the solution using the finite element method based on ten equally spaced nodes. (*iii*) Compare the graphs of the exact solution and its piecewise affine finite element approximation. What is the maximal error in your approximation at the nodes? on the entire interval?

(*a*) $-u'' = \begin{cases} 1 & x > 1, \\ 0, & x < 1, \end{cases}\ u(0) = u(2) = 0$; (*b*) $-\dfrac{d}{dx}\left((1+x)\dfrac{du}{dx}\right) = 1,\ u(0) = u(1) = 0$;

(*c*) $-\dfrac{d}{dx}\left(x^2\dfrac{du}{dx}\right) = -x,\ u(1) = u(3) = 0$; (*d*) $-\dfrac{d}{dx}\left(e^x\dfrac{du}{dx}\right) = e^x,\ u(-1) = u(1) = 0$.

♣ 10.2.3. (*a*) Find the exact solution to the boundary value problem $-u'' = 3x,\ u(0) = u(1) = 0$. (*b*) Use the finite element method based on five equally spaced nodes to approximate the solution. (*c*) Compare the graphs of the exact solution and its piecewise affine finite element approximation. (*d*) What is the maximal error (*i*) at the nodes? (*ii*) on the entire interval?

♣ 10.2.4. Use finite elements to approximate the solution to the Sturm–Liouville boundary value problem $-u'' + (x+1)u = x\,e^x,\ u(0) = 0,\ u(1) = 0$, using 5, 10, and 20 equally spaced nodes.

♣ 10.2.5. (*a*) Devise a finite element scheme for numerically approximating the solution to the mixed boundary value problem

$$-\frac{d}{dx}\left(\kappa(x)\frac{du}{dx}\right) = f(x), \qquad a < x < b, \qquad u(a) = 0, \qquad u'(b) = 0.$$

(*b*) Test your method on the particular boundary value problem

$$-\frac{d}{dx}\left((1+x)\frac{du}{dx}\right) = 1, \qquad 0 < x < 1, \qquad u(0) = 0, \qquad u'(1) = 0,$$

using 10 equally spaced nodes. Compare your approximation with the exact solution.

♠ 10.2.6. Consider the periodic boundary value problem

$$-u'' + u = x, \qquad u(0) = u(2\pi), \qquad u'(0) = u'(2\pi).$$

(*a*) Write down the analytic solution. (*b*) Write down a minimization principle. (*c*) Divide the interval $[0, 2\pi]$ into $n = 5$ equal subintervals, and let W_n denote the subspace consisting of all piecewise affine functions that satisfy the boundary conditions. What is the dimension of W_n? Write down a basis. (*d*) Construct the finite element approximation to the solution to the boundary value problem by minimizing the functional from part (*b*) on the subspace W_n. Graph the result and compare with the exact solution. What is the maximal error on the interval? (*e*) Repeat part (*d*) for $n = 10, 20$, and 40 subintervals, and discuss the convergence of your solutions.

♠ 10.2.7. Answer Exercise 10.2.6 when the finite element subspace W_n consists of all periodic piecewise affine functions of period 1, so $w(x+1) = w(x)$. Which approximation is better?

♣ 10.2.8. Use the method of Exercise 10.2.7 to approximate the solution to the following periodic boundary value problem for the *Mathieu equation*:

$$-u'' + (1 + \cos x)\,u = 1, \qquad u(0) = u(2\pi), \qquad u'(0) = u'(2\pi).$$

♠ 10.2.9. Consider the boundary value problem solved in Example 10.3. Let W_n be the subspace consisting of all polynomials $u(x)$ of degree $\leq n$ satisfying the boundary conditions $u(0) = u(1) = 0$. In this project, we will try to approximate the exact solution to the boundary value problem by minimizing the functional (10.24) on the polynomial subspace W_n. For $n = 5, 10$, and 20: (*a*) First, determine a basis for W_n. (*b*) Set up the minimization problem as a system of linear equations for the coefficients of the polynomial minimizer relative to your basis. (*c*) Solve the polynomial minimization problem and compare your "polynomial finite element" solution with the exact solution and the piecewise affine finite element solution graphed in Figure 10.3.

♠ 10.2.10. Consider the boundary value problem $-u'' + \lambda u = x$, for $0 < x < \pi$, with $u(0) = 0$, $u(1) = 0$. (*a*) For what values of λ does the system have a unique solution? (*b*) For which values of λ can you find a minimization principle that characterizes the solution? Is the minimizer unique for all such values of λ? (*c*) Using n equally spaced nodes, write down the finite element equations for approximating the solution to the boundary value problem. *Note*: Although the finite element construction is supposed to work only when there is a minimization principle, we will consider the resulting linear algebraic system for any value of λ. (*d*) Select a value of λ for which the solution can be characterized by a minimization principle and verify that the finite element approximation with $n = 10$ approximates the exact solution. (*e*) Experiment with other values of λ. Does your finite element solution give a good approximation to the exact solution when it exists? What happens at values of λ for which the solution does not exist or is not unique?

10.3 Finite Elements in Two Dimensions

The same basic framework underlies the adaptation of finite element techniques for numerically approximating the solution to boundary value problems governed by elliptic partial differential equations. In this section, we concentrate on the simplest case: the two-dimensional Poisson equation. Having mastered this, the reader will be well equipped to carry over the method to more general equations and higher dimensions. As before, we concentrate on the practical design of the finite element procedure, and refer the reader to more advanced texts, e.g., [**6**, **113**, **126**], for the analytical details and proofs of convergence. Most of the multi-dimensional complications lie not in the underlying theory, but rather in the realm of data management and organization.

For specificity, consider the homogeneous Dirichlet boundary value problem

$$-\Delta u = f \quad \text{in} \quad \Omega, \qquad\qquad u = 0 \quad \text{on} \quad \partial\Omega, \tag{10.25}$$

on a bounded domain $\Omega \subset \mathbb{R}^2$. According to Theorem 9.31, the solution $u_\star(x, y)$ is characterized as the unique minimizer of the Dirichlet functional

$$Q[u] = \tfrac{1}{2} \,|||\, \nabla u \,|||^2 - \langle\, u \,,\, f \,\rangle = \iint_\Omega \left(\tfrac{1}{2}\, u_x^2 + \tfrac{1}{2}\, u_y^2 - f\, u \right) dx\, dy \tag{10.26}$$

among all C^2 functions $u(x, y)$ that satisfy the prescribed boundary conditions.

To construct a finite element approximation, we restrict the Dirichlet functional to a suitably chosen finite-dimensional subspace. As in the one-dimensional version, the most effective subspaces contain functions that may lack the requisite degree of smoothness that qualifies them as candidate solutions to the partial differential equation. Nevertheless, they will provide good approximations to the actual classical solution. Another important practical consideration, ensuring sparseness of the finite element matrix, is to employ functions

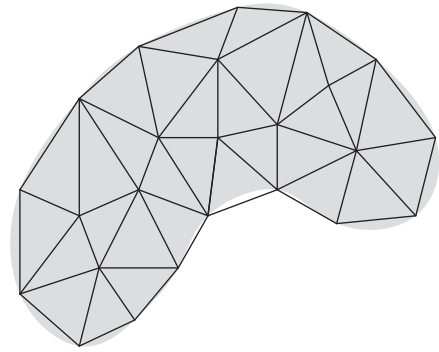


Figure 10.4. Triangulation of a planar domain.

that have small support, meaning that they vanish on most of the domain. Sparseness has the benefit that the solution to the linear finite element system can be relatively rapidly calculated, usually by application of an iterative numerical scheme such as the Gauss–Seidel or SOR methods discussed in [**89**, **118**].

Triangulation

The first step is to introduce a *mesh* consisting of a finite number of *nodes* $\mathbf{x}_l = (x_l, y_l)$, $l = 1, \dots, m$, usually lying inside the domain $\Omega \subset \mathbb{R}^2$. Unlike finite difference schemes, finite element methods are not tied to a rectangular mesh, thus endowing them with considerably more flexibility in the allowable discretizations of the domain. We regard the nodes as the vertices of a *triangulation* of the domain, consisting of a collection of non-overlapping small triangles, which we denote by $T_1, \dots, T_N$, whose union $T_\star = \bigcup_\nu T_\nu$ approximates Ω; see Figure 10.4 for a typical example. The nodes are split into two categories — *interior nodes* and *boundary nodes*, the latter lying on or close to $\partial\Omega$. A curved boundary will thus be approximated by the polygonal boundary $\partial T_\star$ of the triangulation, whose vertexvertices are the boundary nodes. Thus, in any practical implementation of a finite element scheme, the first requirement is a routine that will automatically triangulate a specified domain in some "reasonable" manner, as explained below.

As in our one-dimensional construction, the functions $w(x, y)$ in the finite-dimensional subspace W will be continuous and *piecewise affine*, which means that, on each triangle, the graph of w is a flat plane and hence has the formula[†]

$$w(x, y) = \alpha^\nu + \beta^\nu x + \gamma^\nu y \qquad \text{when} \qquad (x, y) \in T_\nu, \tag{10.27}$$

for certain constants $\alpha^\nu, \beta^\nu, \gamma^\nu$. Continuity of w requires that its values on a common edge between two triangles must agree, and this will impose compatibility constraints on the coefficients $\alpha^\mu, \beta^\mu, \gamma^\mu$ and $\alpha^\nu, \beta^\nu, \gamma^\nu$ associated with adjacent pairs of triangles T_μ and T_ν. The full graph of the piecewise affine function $z = w(x, y)$ forms a connected polyhedral surface whose triangular faces lie above the triangles T_ν; see Figure 10.5 for an illustration. In addition, we require that the piecewise affine function $w(x, y)$ vanish at the boundary nodes, which implies that it vanishes on the entire polygonal boundary of the triangulation,

[†] Here and subsequently, the index ν is a superscript, not a power.

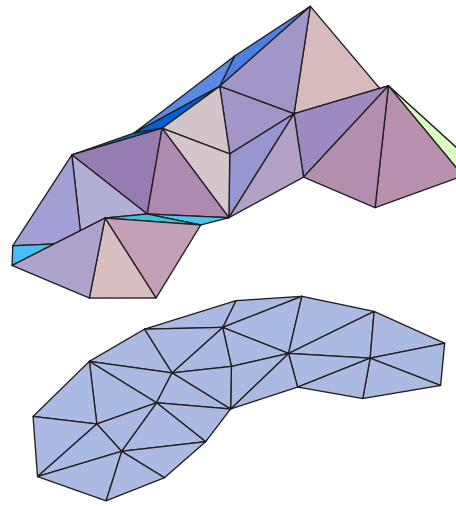


Figure 10.5. Piecewise affine function.

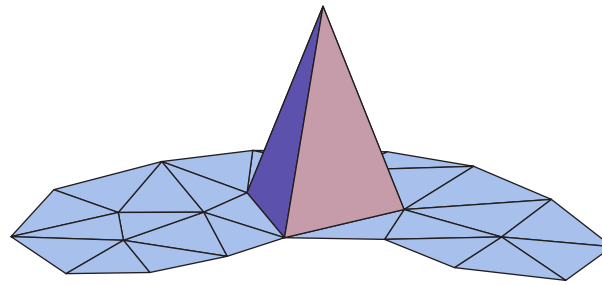


Figure 10.6. Finite element pyramid function.

$\partial T_\star$, and hence (approximately) satisfies the homogeneous Dirichlet boundary conditions on the curved boundary of the original domain, $\partial\Omega$.

The next step is to choose a basis of the subspace of piecewise affine functions associated with the given triangulation and subject to the imposed homogeneous Dirichlet boundary conditions. The analogue of the one-dimensional hat function (10.12) is the *pyramid function* $\varphi_l(x,y)$, which has the value 1 at a single node $\mathbf{x}_l = (x_l, y_l)$, and vanishes at all the other nodes:

$$\varphi_l(x_i, y_i) = \begin{cases} 1, & i = l, \\ 0, & i \neq l. \end{cases} \tag{10.28}$$

Because, on any triangle, the pyramid function $\varphi_l(x,y)$ is uniquely determined by its values at the vertices, it will be nonzero only on those triangles that have the node $\mathbf{x}_l$ as one of their vertices. Hence, as its name implies, the graph of φ_l forms a pyramid of unit height sitting on a flat plane; a typical example appears in Figure 10.6.

The pyramid functions $\varphi_l(x,y)$ associated with the *interior nodes* $\mathbf{x}_l$ automatically satisfy the homogeneous Dirichlet boundary conditions on the boundary of the domain — or, more correctly, on the polygonal boundary of the triangulated domain. Thus, the finite

element subspace W is the span of the interior node pyramid functions, and so a general piecewise affine function $w \in W$ is a linear combination thereof:

$$w(x,y) = \sum_{l=1}^{n} c_l\, \varphi_l(x,y), \tag{10.29}$$

where the sum ranges over the n interior nodes of the triangulation. Owing to the original specification (10.28) of the pyramid functions, the coefficients

$$c_l = w(x_l, y_l) \;\approx\; u(x_l, y_l), \qquad\qquad l = 1, \dots, n, \tag{10.30}$$

are the *same* as the values of the finite element approximation $w(x,y)$ at the interior nodes. This immediately implies linear independence of the pyramid functions, since the only linear combination that vanishes at all nodes is the trivial one $c_1 = \cdots = c_n = 0$.

Determining the explicit formulas for the pyramid functions is not difficult. On one of the triangles T_ν that has $\mathbf{x}_l$ as a vertex, $\varphi_l(x,y)$ will be the unique affine function (10.27) that takes the value 1 at the vertex $\mathbf{x}_l$ and 0 at its other two vertices $\mathbf{x}_i$ and $\mathbf{x}_j$. Thus, we seek a formula for an affine function or *element*

$$\omega_l^\nu(x,y) = \alpha_l^\nu + \beta_l^\nu\, x + \gamma_l^\nu\, y, \qquad\qquad (x,y) \in T_\nu, \tag{10.31}$$

that takes the prescribed values

$$\begin{aligned} \omega_l^\nu(x_i,y_i) &= \alpha_l^\nu + \beta_l^\nu\, x_i + \gamma_l^\nu\, y_i = 0, \\ \omega_l^\nu(x_j,y_j) &= \alpha_l^\nu + \beta_l^\nu\, x_j + \gamma_l^\nu\, y_j = 0, \\ \omega_l^\nu(x_l,y_k) &= \alpha_l^\nu + \beta_l^\nu\, x_l + \gamma_l^\nu\, y_l = 1. \end{aligned} \tag{10.32}$$

Solving this linear system for the coefficients — using either Cramer's Rule or direct Gaussian Elimination — produces the explicit formulas

$$\alpha_l^\nu = \frac{x_i y_j - x_j y_i}{\Delta_\nu}\,, \qquad\qquad \beta_l^\nu = \frac{y_i - y_j}{\Delta_\nu}\,, \qquad\qquad \gamma_l^\nu = \frac{x_j - x_i}{\Delta_\nu}\,, \tag{10.33}$$

where the denominator

$$\Delta_\nu = \det \begin{pmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_l & y_l \end{pmatrix} = \pm 2 \,\text{area}\, T_\nu \tag{10.34}$$

is, up to sign, twice the area of the triangle T_ν; see Exercise 10.3.5.

Example 10.4. Consider an isosceles right triangle T with vertices

$$\mathbf{x}_1 = (0,0), \qquad\qquad \mathbf{x}_2 = (1,0), \qquad\qquad \mathbf{x}_3 = (0,1).$$

Using (10.33–34) (or solving the linear system (10.32) directly), we immediately produce the three corresponding affine elements

$$\omega_1(x,y) = 1 - x - y, \qquad\qquad \omega_2(x,y) = x, \qquad\qquad \omega_3(x,y) = y. \tag{10.35}$$

As required, each ω_l equals 1 at the vertex $\mathbf{x}_l$ and is zero at the other two vertices.

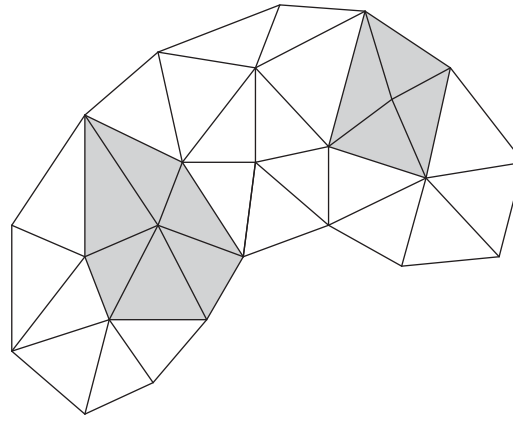


Figure 10.7. Vertex polygons.

A pyramid function is then obtained by piecing together the individual affine elements:

$$\varphi_l(x,y) = \begin{cases} \omega_l^{\nu}(x,y), & \text{if } (x,y) \in T_{\nu} \text{ and } \mathbf{x}_l \text{ is a vertex of } T_{\nu}, \\ 0, & \text{otherwise.} \end{cases} \tag{10.36}$$

Continuity of $\varphi_l(x,y)$ is assured, since the constituent affine elements have the same values at common vertices, and hence also along common edges. The support of the pyramid function (10.36) is the *vertex polygon*

$$\operatorname{supp} \varphi_l = P_l = \bigcup_{\nu} T_{\nu} \tag{10.37}$$

consisting of all the triangles T_{ν} that have the node $\mathbf{x}_l$ as a vertex. In other words, $\varphi_l(x,y) = 0$ whenever $(x,y) \notin P_l$. The node $\mathbf{x}_l$ lies on the interior of its vertex polygon P_l, while the vertices of P_l are all the nodes connected to $\mathbf{x}_l$ by a single edge of the triangulation. In Figure 10.7, the shaded regions indicate two of the vertex polygons for the triangulation in Figure 10.4.

Example 10.5. The simplest, and most common, triangulations are based on regular meshes. For example, suppose that the nodes lie on a square grid, and so are of the form $\mathbf{x}_{i,j} = (i\,h + a, j\,h + b)$, where (i,j) run over a collection of integer pairs, $h > 0$ is the inter-node spacing, and (a,b) represents an overall offset. If we choose the triangles to all have the same orientation, as in the first picture in Figure 10.8, then the vertex polygons all have the same shape, consisting of six triangles of total area $3\,h^2$ — the shaded region. On the other hand, if we choose an alternating triangulation, as in the second picture, then there are two types of vertex polygons. The first, consisting of four triangles, has area $2\,h^2$, while the second, containing eight triangles, has twice the area, $4\,h^2$. In practice, there are good reasons to prefer the former triangulation.

In general, to ensure convergence of the finite element solution to the true minimizer, one should choose triangulations that satisfy the following properties:

- The three side lengths of any individual triangle should be of comparable size, and so long, skinny triangles and obtuse triangles should be avoided.
- The areas of nearby triangles T_{ν} should not vary too much.
- The areas of nearby vertex polygons P_l should also not vary too much.

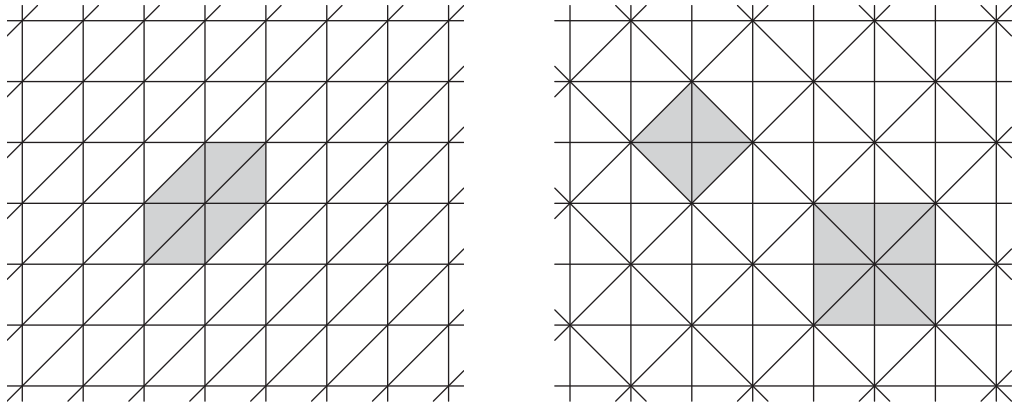


Figure 10.8. Square mesh triangulations.

While the nearby triangles should be of comparable size, one might very well allow wide variations over the entire domain, with small triangles in regions where the solution is changing rapidly, and large triangles in less active regions.

Exercises

10.3.1. Sketch a triangulation of the following domains so that all triangles have side length at most .5: (*a*) a unit square; (*b*) an isosceles triangle with vertices $(-.5, 0), (.5, 0)$ and $(0, 1)$; (*c*) the square $\{ |x|, |y| \leq 2 \}$ with the hole $\{ |x|, |y| < 1 \}$ removed; (*d*) the unit disk; (*e*) the annulus $1 \leq \| \mathbf{x} \| \leq 2$.

10.3.2. Describe the vertex polygons for a triangulation that uses regular equilateral triangles.

10.3.3. Are there any restrictions on the number of sides a vertex polygon can have?

10.3.4. Find the three finite element functions $\omega_1(x, y)$, $\omega_2(x, y)$, $\omega_3(x, y)$, associated with (*a*) the triangle having vertices $(1, 0), (0, 1)$, and $(1, 1)$; (*b*) the triangle having vertices $(0, 1), (1, -1)$, and $(-1, -1)$; (*c*) an equilateral triangle centered at the origin having one vertex at $(1, 0)$.

♢ 10.3.5. (*a*) Prove that the area of a planar triangle T with vertices $(a, b), (c, d), (e, f)$ is equal to $\frac{1}{2} |\Delta|$, where $\Delta = \det \begin{pmatrix} 1 & a & b \\ 1 & c & d \\ 1 & e & f \end{pmatrix}$. (*b*) Prove that $\Delta > 0$ if and only if the vertices of the triangle are listed in counterclockwise order.

♢ 10.3.6. Give a detailed justification of the continuity of the pyramid function (10.36).

♡ 10.3.7. An alternative to triangular elements is to employ piecewise *bi-affine functions*, meaning $\omega(x, y) = \alpha + \beta x + \gamma y + \delta xy$, on rectangles. (*a*) Suppose R is a rectangle with vertices $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$, whose sides are parallel to the coordinate axes. Prove that, for each $l = 1, \ldots, 4$, there is a unique bi-affine function $\omega_l(x, y)$ defined on R that has the value $\omega_l(x_l, y_l) = 1$ at one vertex while $\omega_l(x_i, y_i) = 0$, $i \neq l$, at the other three vertices.

(*b*) Write out the four bi-affine functions $\omega_1(x,y),\dots,\omega_4(x,y)$, when (*i*) $R = \{\, 0 \le x,y \le 1 \,\}$, (*ii*) $R = \{\, -1 \le x,y \le 1 \,\}$. (*c*) Does the result in part (*a*) hold for rectangles whose sides are not aligned with the axes? For general quadrilaterals?

The Finite Element Equations

We now seek to approximate the solution to the homogeneous Dirichlet boundary value problem by restricting the Dirichlet functional (10.26) to the selected finite element subspace W. Using the general framework of Section 10.1, we substitute the formula (10.29) for a general element of W into the quadratic Dirichlet functional (9.82). Expanding, we obtain

$$\begin{aligned} Q[w] = Q\left[\sum_{i=1}^n c_i\,\varphi_i\right] &= \iint_\Omega \left[\left(\sum_{i=1}^n c_i\,\nabla\varphi_i\right)^2 - f(x,y)\,\left(\sum_{i=1}^n c_i\,\varphi_i\right)\right] dx\,dy \\ &= \frac{1}{2}\sum_{i,j=1}^n k_{ij}\,c_i\,c_j - \sum_{i=1}^n b_i\,c_i = \tfrac{1}{2}\,\mathbf{c}^T K\,\mathbf{c} - \mathbf{b}^T\mathbf{c}. \end{aligned} \tag{10.38}$$

Here $K = (k_{ij})$ is a symmetric $n \times n$ matrix, while $\mathbf{b} = (\,b_1, b_2, \dots, b_n\,)^T$ is a vector in $\mathbb{R}^n$, with respective entries

$$\begin{aligned} k_{ij} &= \langle\!\langle\, \nabla\varphi_i\,, \nabla\varphi_j \,\rangle\!\rangle = \iint_\Omega \nabla\varphi_i\cdot\nabla\varphi_j\,dx\,dy, \\ b_i &= \langle\, f\,, \varphi_i\,\rangle = \iint_\Omega f\,\varphi_i\,dx\,dy, \end{aligned} \tag{10.39}$$

which also follow directly from the general formulas (10.6–7). Thus, the finite element approximation (10.29) will minimize the quadratic function

$$P(\mathbf{c}) = \tfrac{1}{2}\,\mathbf{c}^T K\,\mathbf{c} - \mathbf{b}^T\mathbf{c} \tag{10.40}$$

over all possible choices of coefficients $\mathbf{c} = (\,c_1, c_2, \dots, c_n\,)^T \in \mathbb{R}^n$, i.e., over all possible function values at the interior nodes. As above, the minimizer's coefficients are obtained by solving the associated linear system

$$K\,\mathbf{c} = \mathbf{b}, \tag{10.41}$$

using either Gaussian Elimination or a suitable iterative linear systems solver.

To find explicit formulas for the matrix coefficients k_{ij} in (10.39), we begin by noting that the gradient of the affine element (10.31) is equal to

$$\mathbf{g}_l^\nu = \nabla\omega_l^\nu(x,y) = \begin{pmatrix} \partial\omega_l^\nu/\partial x \\ \partial\omega_l^\nu/\partial y \end{pmatrix} = \begin{pmatrix} \beta_l^\nu \\ \gamma_l^\nu \end{pmatrix} = \frac{1}{\Delta_\nu}\begin{pmatrix} y_i - y_j \\ x_j - x_i \end{pmatrix}, \qquad (x,y) \in T_\nu, \tag{10.42}$$

which is a constant vector inside the triangle T_ν, while $\nabla\omega_l^\nu = \mathbf{0}$ outside T_ν. Therefore,

$$\nabla\varphi_l(x,y) = \begin{cases} \mathbf{g}_l^\nu, & \text{if } (x,y) \in T_\nu \text{ that has } \mathbf{x}_l \text{ as a vertex,} \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{10.43}$$

Actually, (10.43) is not quite correct, since the gradient is not well defined on the boundary of a triangle T_ν, but this will not cause us any difficulty in evaluating the ensuing integrals.

We will approximate integrals over the domain Ω by summing the corresponding integrals over the individual triangles — which relies on our assumption that the polygonal boundary of the triangulation $\partial T_\star$ is a reasonably close approximation to the true boundary $\partial\Omega$. In particular,

$$k_{ij} \approx \sum_\nu \iint_{T_\nu} \nabla\varphi_i \cdot \nabla\varphi_j \, dx\, dy \equiv \sum_\nu k_{ij}^\nu. \tag{10.44}$$

Now, according to (10.43), one or the other gradient in the integrand will vanish on the entire triangle T_ν *unless* both $\mathbf{x}_i$ and $\mathbf{x}_j$ are vertices. Therefore, the only terms contributing to the sum are those triangles T_ν that have both $\mathbf{x}_i$ and $\mathbf{x}_j$ as vertices. If $i \neq j$, there are only two such triangles, having a common edge, while if $i = j$, every triangle in the i^{th} vertex polygon P_i contributes. The individual summands are easily evaluated, since the gradients are constant on the triangles, and so, by (10.43),

$$k_{ij}^\nu = \iint_{T_\nu} \mathbf{g}_i^\nu \cdot \mathbf{g}_j^\nu \, dx\, dy = \mathbf{g}_i^\nu \cdot \mathbf{g}_j^\nu \text{ area } T_\nu = \tfrac{1}{2}\, \mathbf{g}_i^\nu \cdot \mathbf{g}_j^\nu \, |\, \Delta_\nu \,|\,.$$

Let T_ν have vertices $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_l$. Then, by (10.34, 42, 43),

$$\begin{aligned} k_{ij}^\nu &= \frac{1}{2}\, \frac{(y_j - y_l)(y_l - y_i) + (x_l - x_j)(x_i - x_l)}{(\Delta_\nu)^2}\, |\, \Delta_\nu \,| = -\, \frac{(\mathbf{x}_i - \mathbf{x}_l)\cdot(\mathbf{x}_j - \mathbf{x}_l)}{2\,|\,\Delta_\nu\,|}, \quad i \neq j, \\ k_{ii}^\nu &= \frac{1}{2}\, \frac{(y_j - y_l)^2 + (x_l - x_j)^2}{(\Delta_\nu)^2}\, |\, \Delta_\nu \,| = \frac{\|\, \mathbf{x}_j - \mathbf{x}_l \,\|^2}{2\,|\,\Delta_\nu\,|} \\ &= -\, \frac{(\mathbf{x}_i - \mathbf{x}_l)\cdot(\mathbf{x}_i - \mathbf{x}_j) + (\mathbf{x}_i - \mathbf{x}_l)\cdot(\mathbf{x}_j - \mathbf{x}_l)}{2\,\Delta_\nu} = -\, k_{ij}^\nu - k_{il}^\nu\,. \end{aligned} \tag{10.45}$$

In this manner, each triangle T_ν specifies a collection of six different coefficients, $k_{ij}^\nu = k_{ji}^\nu$, indexed by its vertices, and known as the *elemental stiffnesses* of T_ν. Interestingly, the elemental stiffnesses depend only on the three vertex *angles* in the triangle and not on its size. Thus, similar triangles have the *same* elemental stiffnesses. Indeed, according to Exercise 10.3.13,

$$k_{ii}^\nu = \tfrac{1}{2}(\cot\theta_j^\nu + \cot\theta_l^\nu), \qquad \text{while} \qquad k_{ij}^\nu = k_{ji}^\nu = -\tfrac{1}{2}\cot\theta_l^\nu, \quad i \neq j, \tag{10.46}$$

where $0 < \theta_l^\nu < \pi$ denotes the angle in T_ν at the vertex $\mathbf{x}_l$.

Example 10.6. The right triangle with vertices $\mathbf{x}_1 = (0,0)$, $\mathbf{x}_2 = (1,0)$, $\mathbf{x}_3 = (0,1)$ has elemental stiffnesses

$$k_{11} = 1, \quad k_{22} = k_{33} = \tfrac{1}{2}, \quad k_{12} = k_{21} = k_{13} = k_{31} = -\tfrac{1}{2}, \quad k_{23} = k_{32} = 0. \tag{10.47}$$

The same holds for any other isosceles right triangle, provided its vertices are labeled in the same manner. Similarly, an equilateral triangle has all 60° angles, and so its elemental stiffnesses are

$$\begin{aligned} &k_{11} = k_{22} = k_{33} = \tfrac{1}{\sqrt{3}} \approx .5774, \\ &k_{12} = k_{21} = k_{13} = k_{31} = k_{23} = k_{32} = -\tfrac{1}{2\sqrt{3}} \approx -.2887. \end{aligned} \tag{10.48}$$

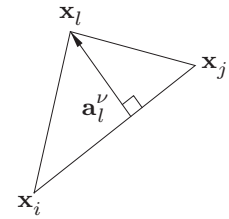
Exercises

10.3.8. Write down the elemental stiffnesses for: (*a*) the triangle with vertices $(0,1)$, $(-1,2)$, $(0,-1)$; (*b*) the triangle with vertices $(1,1)$, $(-1,1)$, $(0,-2)$; (*c*) a $30-60-90$ degree right triangle; (*d*) a right triangle with side lengths $3, 4, 5$; (*e*) an isosceles triangle of height 3 and base 2; (*f*) a "golden" isosceles triangle with angles $36^\circ, 72^\circ, 72^\circ$.

♢ 10.3.9. A *rectangular mesh* has nodes $\mathbf{x}_{i,j} = (i\,\Delta x + a, j\,\Delta y + b)$, where $\Delta x, \Delta y > 0$ are, respectively, the horizontal and vertical step sizes. Find the elemental stiffnesses for the triangles associated with such a rectangular mesh.

10.3.10. *True or false*: Let T be a triangle, and $\widetilde{T}$ a triangle obtained by rotating T by 60°. Then T and $\widetilde{T}$ have the same elemental stiffnesses.

10.3.11. Prove that the gradient (10.42) of the affine element is equal to $\nabla \omega_l^\nu = \| \mathbf{a}_l^\nu \|^{-2} \mathbf{a}_l^\nu$, where $\mathbf{a}_l^\nu$ is the *altitude vector* that goes to the vertex $\mathbf{x}_l$ from its opposite side, as indicated in the figure.



10.3.12. Explain why the pyramid functions are linearly independent.

♢ 10.3.13. Prove formulas (10.46).

Assembling the Elements

The elemental stiffnesses of each triangle will contribute, through the summation (10.44), to the finite element coefficient matrix K. We begin by constructing a larger matrix $\widehat{K}$, which we call the *full finite element matrix*, of size $m \times m$, where m is the total number of nodes in our triangulation, including both interior and boundary nodes. The rows and columns of $\widehat{K}$ are labeled by the nodes $\mathbf{x}_1, \dots, \mathbf{x}_m$. Let $K_\nu = (k_{ij}^\nu)$ be the corresponding $m \times m$ matrix containing the elemental stiffnesses k_{ij}^ν of T_ν in the rows and columns indexed by its vertices, and all other entries equal to 0. Thus, K_ν will have (at most) nine nonzero entries. The resulting $m \times m$ matrices are summed together over all the triangles $T_1, \dots, T_N$, whereby

$$\widehat{K} = \sum_{\nu=1}^{N} K_\nu, \tag{10.49}$$

in accordance with (10.44).

The full finite element matrix $\widehat{K}$ is too large, since its rows and columns include all the nodes, whereas the finite element matrix K appearing in (10.41) refers only to the n interior nodes. The *reduced* $n \times n$ *finite element matrix* K is simply obtained from $\widehat{K}$ by deleting all rows and columns indexed by boundary nodes, retaining only the elements k_{ij} for which both $\mathbf{x}_i$ and $\mathbf{x}_j$ are interior nodes. For the homogeneous boundary value problem, this is all we require. As we will subsequently see, inhomogeneous boundary conditions are most easily handled by retaining (another part of) the full matrix $\widehat{K}$.

The easiest way to absorb the construction is by working through a particular example.

Example 10.7. A metal plate has the shape of an oval running track, consisting of a rectangle, with side lengths 1 m by 2 m, and two semi-circular disks glued onto its

Figure 10.9. The oval plate.

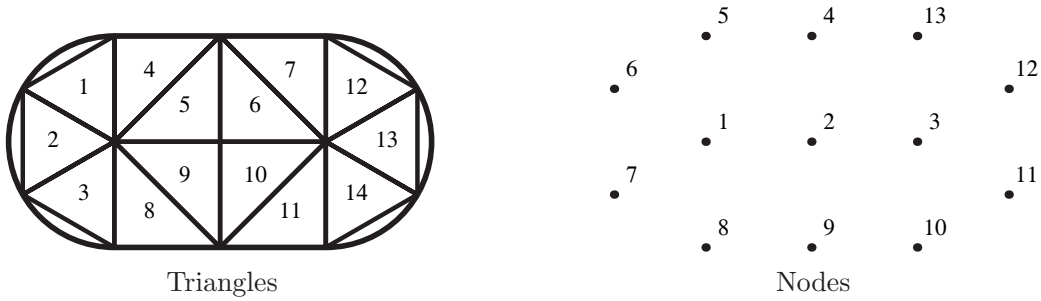


Figure 10.10. A coarse triangulation of the oval plate.

shorter ends, as sketched in Figure 10.9. The plate is subject to a heat source, while its edges are held at a fixed temperature. The problem is to find the equilibrium temperature distribution within the plate. Mathematically, we must solve the planar Poisson equation, subject to Dirichlet boundary conditions, for the equilibrium temperature $u(x, y)$.

Let us describe how to set up the finite element approximation. We begin with a very coarse triangulation of the plate, which will not give particularly accurate results, but serves to illustrate how to go about assembling the finite element matrix. We divide the rectangular part of the plate into eight right triangles, while each semicircular end will be approximated by three equilateral triangles. The triangles are numbered from 1 to 14 as indicated in Figure 10.10. There are 13 nodes in all, numbered as in the second figure. Only nodes $1, 2, 3$ are interior, while the boundary nodes are labeled 4 through 13 in counterclockwise order starting at the top. The full finite element matrix $\widehat{K}$ will have size 13×13, its rows and columns labeled by all the nodes, while the reduced matrix K appearing in the finite element equations (10.41) consists of the upper left 3×3 submatrix of $\widehat{K}$ corresponding to the three interior nodes.

For each $\nu = 1, \dots, 14$, the triangle T_ν will contribute its elemental stiffnesses, as indexed by its vertices, to the matrix $\widehat{K}$ through a summand K_ν. For example, the first triangle T_1 is equilateral, and so has elemental stiffnesses (10.48). Its vertices are labeled 1, 5, and 6, and therefore we place the stiffnesses in the rows and columns numbered $1, 5, 6$

to form the summand

$$K_1 = \begin{pmatrix} .5774 & 0 & 0 & 0 & -.2887 & -.2887 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ -.2887 & 0 & 0 & 0 & .5774 & -.2887 & 0 & 0 & \dots \\ -.2887 & 0 & 0 & 0 & -.2887 & .5774 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where all the undisplayed entries in the full 13×13 matrix are 0. The next triangle T_2 has the same equilateral elemental stiffness matrix (10.48), but now its vertices are $1, 6, 7$, and so it will contribute

$$K_2 = \begin{pmatrix} .5774 & 0 & 0 & 0 & 0 & -.2887 & -.2887 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ -.2887 & 0 & 0 & 0 & 0 & .5774 & -.2887 & 0 & \dots \\ -.2887 & 0 & 0 & 0 & 0 & -.2887 & .5774 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Similarly for K_3, with vertices $1, 7, 8$. On the other hand, T_4 is an isosceles right triangle, and so has elemental stiffnesses (10.47). Its vertices are labeled 1, 4, and 5, with vertex 5 at the right angle. Therefore, its contribution is

$$K_4 = \begin{pmatrix} .5 & 0 & 0 & 0 & -.5 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & .5 & -.5 & 0 & 0 & 0 & \dots \\ -.5 & 0 & 0 & -.5 & 1.0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Continuing in this manner, we assemble 14 contributions $K_1, \dots, K_{14}$, each with at most 9 nonzero entries. The full finite element matrix is their sum

$$\widehat{K} = K_1 + K_2 + \cdots + K_{14}$$

$$= \begin{pmatrix} 3.732 & -1 & 0 & 0 & -.7887 & -.5774 & -.5774 \\ -1 & 4 & -1 & -1 & 0 & 0 & 0 \\ 0 & -1 & 3.732 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 2 & -.5 & 0 & 0 \\ -.7887 & 0 & 0 & -.5 & 1.577 & -.2887 & 0 \\ -.5774 & 0 & 0 & 0 & -.2887 & 1.155 & -.2887 \\ -.5774 & 0 & 0 & 0 & 0 & -.2887 & 1.155 \\ -.7887 & 0 & 0 & 0 & 0 & 0 & -.2887 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -.7887 & 0 & 0 & 0 & 0 \\ 0 & 0 & -.5774 & 0 & 0 & 0 & 0 \\ 0 & 0 & -.5774 & 0 & 0 & 0 & 0 \\ 0 & 0 & -.7887 & -.5 & 0 & 0 & 0 \end{pmatrix} \tag{10.50}$$

$$\begin{pmatrix} -.7887 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -.7887 & -.5774 & -.5774 & -.7887 \\ 0 & 0 & 0 & 0 & 0 & -.5 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -.2887 & 0 & 0 & 0 & 0 & 0 \\ 1.577 & -.5 & 0 & 0 & 0 & 0 \\ -.5 & 2 & -.5 & 0 & 0 & 0 \\ 0 & -.5 & 1.577 & -.2887 & 0 & 0 \\ 0 & 0 & -.2887 & 1.155 & -.2887 & 0 \\ 0 & 0 & 0 & -.2887 & 1.155 & -.2887 \\ 0 & 0 & 0 & 0 & -.2887 & 1.577 \end{pmatrix}.$$

Since nodes $1, 2, 3$ are interior, the reduced finite element matrix

$$K = \begin{pmatrix} 3.732 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 3.732 \end{pmatrix} \tag{10.51}$$

uses only the upper left 3×3 block of $\widehat{K}$. Clearly, it would not be difficult to directly construct K, bypassing $\widehat{K}$ entirely.

For a finer triangulation, the construction is similar, but the matrices become much larger. The procedure can, of course, be automated. Fortunately, if we choose a very regular triangulation, then we do not need to be nearly as meticulous in assembling the stiffness matrices, since many of the entries are the same. The simplest case employs a uniform square mesh, and so triangulates the domain into isosceles right triangles. This is accomplished by laying out a relatively dense square grid over the domain $\Omega \subset \mathbb{R}^2$. The interior nodes are the grid points that fall inside the oval domain, while the boundary nodes are all those grid points lying adjacent to one or more of the interior nodes, and are near but not necessarily precisely on the boundary $\partial\Omega$. Figure 10.11 shows the nodes in a square grid with intermesh spacing $h = .2$. While a bit crude in its approximation of the boundary of the domain, this procedure does have the advantage of making the construction of the associated finite element matrix relatively painless.

For such a mesh, all the triangles are isosceles right triangles, with elemental stiffnesses (10.47). Summing the corresponding matrices K_ν over all the triangles, as in (10.49), we find that the rows and columns of $\widehat{K}$ corresponding to the interior nodes all have the same

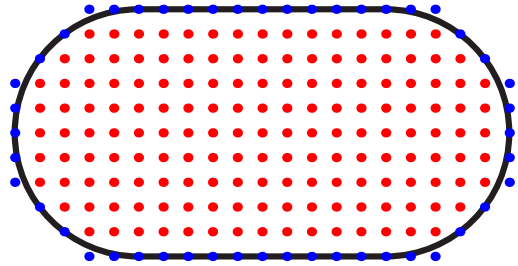


Figure 10.11. A square mesh for the oval plate.

form. Namely, if i labels an interior node, then the corresponding diagonal entry is $k_{ii} = 4$, while the off-diagonal entries $k_{ij} = k_{ji}$, $i \neq j$, are equal to -1 when node i is adjacent to node j on the grid, and are equal to 0 in all other cases. Node j is allowed to be a boundary node. (Interestingly, the result does not depend on how one orients the pair of triangles making up each square of the grid, which plays a role only in the computation of the right-hand side of the finite element equation.) Observe that the same computation applies even to our coarse triangulation. The interior node 2 belongs to all right isosceles triangles, and the corresponding nonzero entries in (10.50) are $k_{22} = 4$ and $k_{21} = k_{23} = k_{24} = k_{29} = -1$, indicating the four adjacent nodes.

Remark: The coefficient matrix constructed from the finite element method on a square (or even rectangular) grid is the *same* as the coefficient matrix arising from a finite difference solution to the Laplace or Poisson equation, as described in Example 5.7. The finite element approach has the advantage of readily adapting to much more general discretizations of the domain, and is not restricted to rectangular grids.

The Coefficient Vector and the Boundary Conditions

So far, we have been concentrating on assembling the finite element coefficient matrix K. We also need to compute the forcing vector $\mathbf{b} = (\, b_1, b_2, \ldots, b_n \,)^T$ appearing on the right-hand side of the fundamental linear equation (10.41). According to (10.39), the entries b_i are found by integrating the product of the forcing function and the finite element basis function. As before, we will approximate the integral over the domain Ω by an integral over the triangles, and so

$$b_i = \iint_\Omega f(x,y)\, \varphi_i(x,y)\, dx\, dy \approx \sum_\nu \iint_{T_\nu} f(x,y)\, \omega_i^\nu(x,y)\, dx\, dy \equiv \sum_\nu b_i^\nu. \qquad (10.52)$$

Typically, an exact computation of the various triangular double integrals is not so convenient, and so we resort to a numerical approximation. Since we are assuming that the individual triangles are small, we can get away with a very crude numerical integration scheme. If the function $f(x,y)$ does not vary much over the triangle T_ν — which will certainly be the case if T_ν is sufficiently small — we may approximate $f(x,y) \approx c_i^\nu$ for

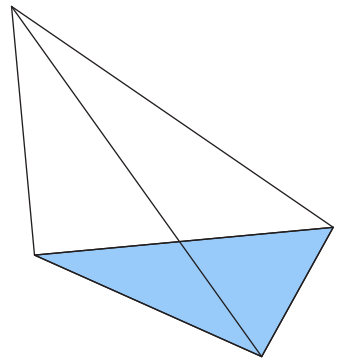


Figure 10.12. Finite element tetrahedron.

$(x, y) \in T_\nu$ by a constant. The integral (10.52) is then approximated by

$$b_i^\nu = \iint_{T_\nu} f(x,y)\, \omega_i^\nu(x,y)\, dx\, dy \approx c_i^\nu \iint_{T_\nu} \omega_i^\nu(x,y)\, dx\, dy = \tfrac{1}{3}\, c_i^\nu \text{ area } T_\nu = \tfrac{1}{6}\, c_i^\nu\, |\, \Delta_\nu\, |. \tag{10.53}$$

The formula for the integral of the affine element $\omega_i^\nu(x, y)$ follows from solid geometry: it equals the volume under its graph, a tetrahedron of height 1 and base T_ν, as illustrated in Figure 10.12.

How to choose the constant c_i^ν? In practice, the simplest choice is to let $c_i^\nu = f(x_i, y_i)$ be the value of the function at the i^{th} vertex. With this choice, the sum in (10.52) becomes

$$b_i \approx \sum_\nu \tfrac{1}{3}\, f(x_i, y_i) \text{ area } T_\nu = \tfrac{1}{3}\, f(x_i, y_i) \text{ area } P_i, \tag{10.54}$$

where P_i is the vertex polygon (10.37) corresponding to the node $\mathbf{x}_i$. In particular, for the square mesh with the uniform choice of triangles, as in the first plot in Figure 10.8,

$$\text{area } P_i = 3\, h^2 \qquad \text{for all } i, \text{ and so} \qquad b_i \approx f(x_i, y_i)\, h^2 \tag{10.55}$$

is well approximated by just h^2 times the value of the forcing function at the node. This is the underlying reason to choose the uniform triangulation for the square mesh; the alternating version would give unequal values for the b_i over adjacent nodes, and this could give rise to unnecessary errors in the final approximation.

Example 10.8. For the coarsely triangulated oval plate, the reduced stiffness matrix is (10.51). The Poisson equation

$$-\Delta u = 4$$

models a constant external heat source of magnitude 4° over the entire plate. If we keep the edges of the plate fixed at 0°, then we need to solve the finite element equation $K\,\mathbf{c} = \mathbf{b}$, where K is the coefficient matrix (10.51). The entries of $\mathbf{b}$ are, by (10.54), equal to 4 (the right-hand side of the differential equation) times one-third the area of the corresponding vertex polygon, which for node 2 is the square consisting of four right triangles, each of area $\frac{1}{2}$, whereas for nodes 1 and 3 it consists of four right triangles of area $\frac{1}{2}$ plus three

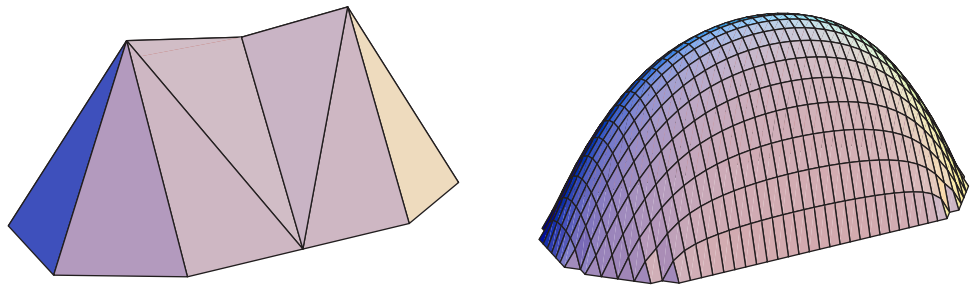


Figure 10.13. Finite element solutions to Poisson's equation for an oval plate.

equilateral triangles, each of area $\frac{\sqrt{3}}{4}$; see Figure 10.10. Thus,

$$\mathbf{b} = \tfrac{4}{3}\left(2 + \tfrac{3\sqrt{3}}{4},\ 2,\ 2 + \tfrac{3\sqrt{3}}{4}\right)^T = (\,4.3987, 2.6667, 4.3987\,)^T .$$

The solution to the final linear system $K\,\mathbf{c} = \mathbf{b}$ is easily found:

$$\mathbf{c} = (\,1.5672, 1.4503, 1.5672\,)^T .$$

Its entries are the values of the finite element approximation at the three interior nodes. The piecewise affine finite element solution is plotted in the first illustration in Figure 10.13. A more accurate approximation, based on a square grid triangulation of size $h = .1$, appears in the second figure. Here, the largest errors are concentrated near the poorly approximated corners of the oval, and could be improved by a more sophisticated triangulation.

Inhomogeneous Boundary Conditions

So far, we have restricted our attention to problems with homogeneous Dirichlet boundary conditions. According to Theorem 9.32, the solution to the inhomogeneous Dirichlet problem

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

is also obtained by minimizing the Dirichlet functional (9.82). However, now the minimization takes place over the set of functions that satisfy the inhomogeneous boundary conditions. It is not difficult to fit this problem into the finite element scheme.

The elements corresponding to the interior nodes of our triangulation remain as before, but now we need to include additional elements to ensure that our approximation satisfies the boundary conditions. Note that if $\mathbf{x}_l$ is a boundary node, then the corresponding *boundary element* $\varphi_l(x, y)$ satisfies (10.28), and so has the same piecewise affine form (10.36). The corresponding finite element approximation

$$w(x, y) = \sum_{l=1}^{m} c_l\, \varphi_l(x, y) \tag{10.56}$$

has the same form as before, (10.29), but now the sum is over *all* nodes, both interior and boundary. As before, the coefficients $c_l = w(x_l, y_l) \approx u(x_l, y_l)$ are the values of the

finite element approximation at the nodes. Therefore, in order to satisfy the boundary conditions, we require

$$c_j = h_j = h(x_j, y_j) \qquad \text{whenever} \quad \mathbf{x}_j = (x_j, y_j) \quad \text{is a boundary node.} \tag{10.57}$$

If the boundary node $\mathbf{x}_j$ does not lie precisely on the boundary $\partial\Omega$, then $h(x_j, y_j)$ is not defined, and so we need to approximate the value h_j appropriately, e.g., using the value of $h(x, y)$ at a nearby boundary point $(x, y) \in \partial\Omega$.

The derivation of the finite element equations proceeds as before, but now there are additional terms arising from the nonzero boundary values. Leaving the intervening details to Exercise 10.3.23, the final outcome can be written as follows. Let $\widehat{K}$ denote the full $m \times m$ finite element matrix constructed as above. The reduced coefficient matrix K is obtained by retaining the rows and columns corresponding to only interior nodes, and so will have size $n \times n$, where n is the number of interior nodes. The *boundary coefficient matrix* $\widetilde{K}$ is the $n \times (m - n)$ matrix consisting of those entries of the interior rows that do not appear in K, i.e., those lying in the columns indexed by the boundary nodes. For instance, in the coarse triangulation of the oval plate, the full finite element matrix is given in (10.50), and the upper 3×3 subblock is the reduced matrix (10.51). The remaining entries of the first three rows form the boundary coefficient matrix

$$\widetilde{K} = \begin{pmatrix} 0 & -.7887 & -.5774 & -.5774 & -.7887 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -.7887 & -.5774 & -.5774 & -.7887 \end{pmatrix}. \tag{10.58}$$

We similarly split the coefficients c_i of the finite element function (10.56) into two groups. We let $\mathbf{c} = (\, c_1, c_2, \ldots, c_n \,)^T \in \mathbb{R}^n$ denote the as yet unknown coefficients corresponding to the values of the approximation at the interior nodes $\mathbf{x}_i$, while $\mathbf{h} = \big(\, h_1, h_2, \ldots, h_{m-n} \,\big)^T \in \mathbb{R}^{m-n}$ will be the vector containing the boundary values (10.57). The solution to the finite element approximation (10.56) is then obtained by solving the associated linear system

$$K\,\mathbf{c} + \widetilde{K}\,\mathbf{h} = \mathbf{b}, \qquad \text{or, equivalently,} \qquad K\,\mathbf{c} = \mathbf{f} = \mathbf{b} - \widetilde{K}\,\mathbf{h}. \tag{10.59}$$

Example 10.9. For the oval plate discussed in Example 10.7, suppose the right-hand semicircular edge is held at $10°$, the left-hand semicircular edge at $-10°$, while the two straight edges have a linearly varying temperature distribution ranging from $-10°$ at the left to $10°$ at the right, as illustrated in Figure 10.14. Our task is to compute its equilibrium temperature, assuming no internal heat source. Thus, for the coarse triangulation we have the boundary node values

$$\mathbf{h} = (\, h_4, \ldots, h_{13} \,)^T = (\, 0, -10, -10, -10, -10, 0, 10, 10, 10, 10 \,)^T.$$

Using the previously computed formulas (10.51, 58) for the interior and boundary coefficient matrices $K, \widetilde{K}$, we approximate the solution to the Laplace equation by solving (10.59). We are assuming that there is no external forcing function, $f(x, y) \equiv 0$, and hence $\mathbf{b} = \mathbf{0}$, and so we must solve $K\,\mathbf{c} = \mathbf{f} = -\widetilde{K}\,\mathbf{h} = (\, 2.1856, 3.6, 7.6497 \,)^T$. The finite element function corresponding to the solution $\mathbf{c} = (\, 1.0679, 1.8, 2.5320 \,)^T$ is plotted in the first illustration in Figure 10.14. Even on such a coarse mesh, the approximation is not too bad, as evidenced by the second illustration, which plots the finite element solution for the finer square mesh of Figure 10.11.

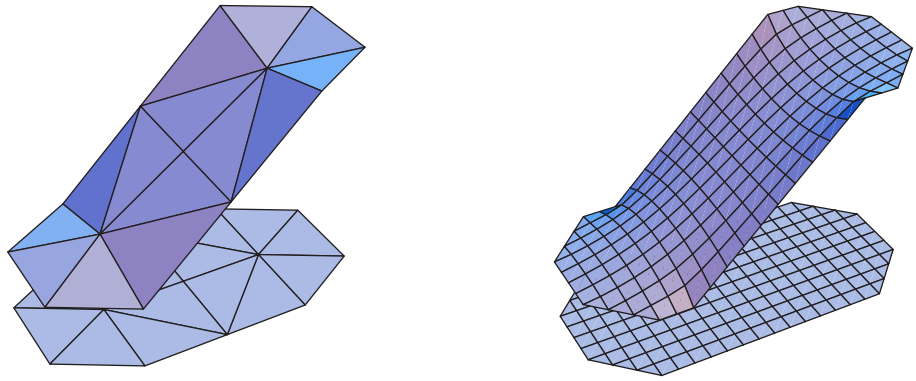


Figure 10.14. Solution to the Dirichlet problem for the oval plate.

Exercises

♠ 10.3.14. Consider the Dirichlet boundary value problem $\Delta u = 0$, $u(x,0) = \sin x$, $u(x,\pi) = 0$, $u(0,y) = 0$, $u(\pi,y) = 0$, on the square $S = \{\, 0 < x, y < \pi \,\}$. (*a*) Find the exact solution. (*b*) Set up and solve the finite element equations based on a square mesh with $n = 2$ squares on each side of S. Write out the reduced finite element matrix, the boundary coefficient matrix, and the value of your approximation at the middle of the unit square. How close is this value to the exact solution there? (*c*) Repeat part (*b*) for $n = 4$ squares per side. Is the value of your approximation at the center of the unit square closer to the true solution? (*d*) Use a computer to find a finite element approximation to $u\left(\frac{1}{2}\pi, \frac{1}{2}\pi\right)$ using $n = 8$ squares per side. Is your approximation converging to the exact solution as the mesh becomes finer and finer?

♣ 10.3.15. Approximate the solution to the Dirichlet problem $\Delta u = 0$, $u(x,0) = x$, $u(x,1) = 1 - x$, $u(0,y) = y$, $u(1,y) = 1 - y$, by use of finite elements with mesh sizes $\Delta x = \Delta y = .25$ and .1. Compare your approximations with the solution you obtained in Exercise 4.3.12(*d*). What is the maximal error at the nodes in each case?

♠ 10.3.16. A metal plate has the shape of an equilateral triangle with unit sides. One side is heated to 100°, while the other two are kept at 0°. In order to approximate the equilibrium temperature distribution, the plate is divided into smaller equilateral triangles, with n triangles on each side, and the corresponding finite element approximation is then computed. (*a*) How many triangles are in the triangulation? How many interior nodes? How many edge nodes? (*b*) For $n = 2$, set up and solve the finite element linear system to find an approximation to the temperature at the center of the triangle. (*c*) Answer part (*b*) when $n = 3$. (*d*) Use a computer to find the finite element approximation to the temperature at the center when $n = 5, 10$, and 15. Are your values converging to the actual temperature? (*e*) Plot the finite element approximations you constructed in the previous parts.

10.3.17. Find the equilibrium temperature distribution in a unit equilateral triangle when one side is heated to 100°, while the other two are insulated.

♠ 10.3.18. A metal plate has the shape of a 3 cm square with a 1 cm square hole cut out of the middle. The plate is heated by fixing the inner edge at temperature 100° while keeping the outer edge at 0°. (*a*) Find the (approximate) equilibrium temperature using finite

elements with a mesh width of $\Delta x = \Delta y = .5$ cm. Plot your approximate solution using a three-dimensional graphics program. (*b*) Let C denote the square contour lying midway between the inner and outer square boundaries of the plate. Using your finite element approximation, at what point(s) on C is the temperature a (*i*) minimum? (*ii*) maximum? (*iii*) equal to 50°, the average of the two boundary temperatures? (*c*) Repeat part (*a*) using a smaller mesh width of $h = .2$. How much does this affect your answers in part (*b*)?

♣ 10.3.19. Answer Exercise 10.3.18 when the plate is additionally subjected to a constant heat source $f(x, y) = 600\,x + 800\,y - 2400$.

♠ 10.3.20. (*a*) Construct a finite element approximation to the solution, using a maximal mesh size of .1, to the following boundary value problem on the unit disk:

$$\Delta u = 0, \qquad x^2 + y^2 < 1, \qquad u = \begin{cases} 1, & x^2 + y^2 = 1, \quad y > 0, \\ 0, & x^2 + y^2 = 1, \quad y < 0. \end{cases}$$

(*b*) Compare your solution with the exact solution given in Example 4.7.

♣ 10.3.21. (*a*) Use finite elements to approximate the solution to the boundary value problem

$$-\Delta u + u = 0, \qquad 0 < x, y < 1, \qquad u(x, 0) = u(x, 1) = u(0, y) = 0, \qquad u(1, y) = 1.$$

(*b*) Compare your result with the first 5 and 10 summands in the series solution obtained via separation of variables.

♢ 10.3.22. (*a*) Justify the construction of the finite element matrix for a square mesh described in the text. (*b*) How would you modify the matrix for a rectangular mesh, as in Exercise 10.3.9?

♢ 10.3.23. Justify the inhomogeneous finite element construction in the text.

♡ 10.3.24. (*a*) Explain how to adapt the finite element method to a mixed boundary value problem with inhomogeneous Neumann conditions. (*b*) Apply your method to the problem

$$\Delta u = 0, \qquad \frac{\partial u}{\partial y}\,(x, 0) = x, \qquad u(x, 1) = 0, \qquad u(0, y) = 0, \qquad u(1, y) = 0.$$

(*c*) Solve the boundary value problem via separation of variables. Compare the values of your solutions at the center of the square.

10.4 Weak Solutions

An alternative route to the finite element method, which avoids the requirement of a minimization principle, rests upon the notion of a weak solution to a differential equation — a concept of considerable independent interest, since it includes many of the nonclassical solutions that we encountered earlier in this book. In particular, the discontinuous shock waves of Section 2.3 are, in fact, weak solutions to the nonlinear transport equation, as are the continuous but only piecewise smooth solutions to the wave equation that resulted from applying d'Alembert's formula to nonsmooth initial data. Weak solutions have become an incredibly powerful idea in the modern theory of partial differential equations, and we have space to present only the very basics here. They are particularly appropriate in the study of discontinuous and nonsmooth physical phenomena, including shock waves, cracks and dislocations in elastic media, singularities in liquid crystals, and so on. In the mathematical analysis of partial differential equations, it is often easier to prove the existence of a weak solution, for which one can then try to establish sufficient smoothness in order that it qualify as a classical solution. Further developments along with a range of applications can be found in more advanced texts, including [**38**, **44**, **61**, **99**, **107**, **122**].

Weak Formulations of Linear Systems

The key idea behind the concept of a weak solution begins with a rather trivial observation: the only element in an inner product space that is orthogonal to every other element is the zero element.

Lemma 10.10. *Let V be an inner product space with inner product*† $\langle\!\langle \cdot , \cdot \rangle\!\rangle$. *An element $v_\star \in V$ satisfies $\langle\!\langle v_\star , v \rangle\!\rangle = 0$ for all $v \in V$ if and only if $v_\star = 0$.*

Proof: In particular, $v_\star$ must be orthogonal to itself, so $0 = \langle\!\langle v_\star , v_\star \rangle\!\rangle = |||\, v_\star \,|||^2$, which immediately implies $v_\star = 0$. *Q.E.D.*

Thus, one method of solving a linear — or even nonlinear — equation $F[u] = 0$ is to write it in the form

$$\langle\!\langle F[u] , v \rangle\!\rangle = 0 \qquad \text{for all} \qquad v \in V, \tag{10.60}$$

where V is the target space of $F\colon U \to V$. In particular, for an inhomogeneous linear system, $L[u] = f$, with $L\colon U \to V$ a linear operator between inner product spaces, the condition (10.60) takes the form

$$0 = \langle\!\langle L[u] - f , v \rangle\!\rangle = \langle\!\langle L[u] , v \rangle\!\rangle - \langle\!\langle f , v \rangle\!\rangle \qquad \text{for all} \qquad v \in V,$$

or, equivalently,

$$\langle u , L^*[v] \rangle - \langle\!\langle f , v \rangle\!\rangle = 0 \qquad \text{for all} \qquad v \in V, \tag{10.61}$$

where $L^*\colon V \to U$ denotes the adjoint of the operator L, as defined in (9.2). We will call (10.61) the *weak formulation* of the original linear system.

So far we have not really done anything of substance, and, indeed, for linear systems of algebraic equations, this more complicated characterization of solutions is of scant help. However, this is no longer the case for differential equations, because, thanks to the integration by parts argument used to determine the adjoint operator, the solution u to the weak form (10.61) is not restricted by the degree of smoothness required of a classical solution. A simple example will illustrate the basic construction.

Example 10.11. On a bounded interval $a \le x \le b$, consider the elementary boundary value problem

$$-\frac{d^2u}{dx^2} = f(x), \qquad u(a) = u(b) = 0.$$

The underlying vector space is $U = \{\, u(x) \in \mathrm{C}^2[a,b] \mid u(a) = u(b) = 0 \,\}$. To obtain a weak formulation, we multiply the differential equation by a test function $v(x) \in U$ and integrate:

$$\int_a^b \big[-u''(x) - f(x) \big] v(x)\, dx = 0. \tag{10.62}$$

The left-hand integral can be identified with the L^2 inner product between the left-hand side of the equation $L[u] - f = -u'' - f = 0$ and the test function v. According to Lemma 10.10, condition (10.62) holds for all $v(x) \in U$ if and only if $u(x) \in U$ satisfies the

† Shortly, as in the general framework developed in Chapter 9, V will be identified as the target space of a linear operator $L\colon U \to V$, and hence the choice of notation for its inner product.

boundary value problem. However, suppose that we integrate the first term by parts once. The boundary conditions on v imply that the boundary terms vanish, and the result is

$$\int_a^b \big[\, u'(x)\, v'(x) - f(x)\, v(x) \,\big]\, dx = 0. \tag{10.63}$$

A function $u(x)$ that satisfies the latter integral condition for all smooth test functions $v(x)$ will be called a *weak solution* to the original boundary value problem. The key observation is that the original differential equation, as well as the integral reformulation (10.62), requires that $u(x)$ be twice differentiable, whereas the weak version (10.63) requires only that its first derivative be defined.

Of course, one need not stop at (10.63). Performing another integration by parts on its first term and invoking the boundary conditions on u produces

$$\int_a^b \big[\, -u(x)\, v''(x) - f(x)\, v(x) \,\big]\, dx = 0. \tag{10.64}$$

Now $u(x)$ need only be (piecewise) continuous in order that the integral be defined — keeping in mind that the test function $v(x)$ is still required to be smooth. Equation (10.64) is sometimes referred to as the *fully weak formulation* of the boundary value problem, while the intermediate integral (10.63), in which the derivatives are evenly distributed among u and v, is then known as the *semi-weak formulation*.

Remark: Recall also the Definition 6.5 of weak convergence, which similarly involves integrating the standard convergence criterion against a suitable test function. Both are part and parcel of a general weak analytical framework that plays an essential role in all of modern advanced analysis, including partial differential equations.

The preceding example is a particular case of a general construction based on the abstract formulation of self-adjoint linear systems in Chapter 9. Let $L\colon U \to V$ be a linear map between inner product spaces, and let $S = L^* \circ L : U \to U$ be the associated self-adjoint operator. We further assume that $\ker L = \{0\}$, which implies that $S > 0$ is positive definite and, provided $f \in \operatorname{rng} S$, the associated linear system

$$S[u] = L^* \circ L[u] = f \tag{10.65}$$

has a unique solution.

In order to construct a weak formulation of the linear system (10.65), we begin by taking its inner product with a test function $v \in U$, whereby

$$0 = \langle\, S[u] - f\,, v\,\rangle = \langle\, S[u]\,, v\,\rangle - \langle\, f\,, v\,\rangle = \langle\, L^* \circ L[u]\,, v\,\rangle - \langle\, f\,, v\,\rangle.$$

Integration by parts, as in the preceding example, amounts to moving the adjoint operator so that it acts on the test function v, and in this manner we obtain the *weak formulation*

$$\langle\!\langle\, L[u]\,, L[v]\,\rangle\!\rangle = \langle\, f\,, v\,\rangle \qquad \text{for all} \qquad v \in U, \tag{10.66}$$

where we use our usual notation conventions regarding the inner products on U and V.

Warning: Unlike the minimization principle (10.1), the weak formulation (10.66) does not have a factor of $\frac{1}{2}$ on the left-hand side. Since, in the applications treated here, L is a differential operator of order, say, k, the weak formulation requires only that $u \in \mathrm{C}^k$ be k times differentiable, whereas, since S has order $2\,k$, the classical formulation (10.65) requires $u \in \mathrm{C}^{2k}$ to have twice as many derivatives.

Similarly, the fully weak formulation involves an additional integration by parts, realized in the abstract framework by moving the linear operator L acting on u so as to act on the test element v, and so

$$\langle\, u\,, L^* \circ L[\,v\,]\,\rangle = \langle\, u\,, S[\,v\,]\,\rangle = \langle\, f\,, v\,\rangle \qquad \text{for all} \qquad v \in U. \tag{10.67}$$

In practice, it is often advantageous to restrict the class of test functions in order to avoid technicalities involving smoothness and boundary behavior. This requires replacing the simple argument used to establish Lemma 10.10 by a more sophisticated result, named after the nineteenth-century German analyst Paul du Bois–Reymond.

Lemma 10.12. *Let $f(x)$ be a continuous function for $a \le x \le b$. Then*

$$\int_a^b f(x)\, v(x)\, dx = 0$$

for every C^1 function $v(x)$ with compact support in the open interval (a, b) if and only if $f(x) \equiv 0$.

Proof: Suppose $f(x_0) > 0$ for some $a < x_0 < b$. Then, by continuity, $f(x) > 0$ for all x in some interval $a < x_0 - \varepsilon < x < x_0 + \varepsilon < b$ around x_0. Choose $v(x)$ to be a C^1 function that is strictly positive in this interval and vanishes outside. An example is

$$v(x) = \begin{cases} \big((x - x_0)^2 - \varepsilon^2\big)^2, & |\, x - x_0\,| \le \varepsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{10.68}$$

Then $f(x)\, v(x) > 0$ when $|\, x - x_0\,| < \varepsilon$ and $= 0$ everywhere else. This implies

$$\int_a^b f(x)\, v(x)\, dx = \int_{x_0-\varepsilon}^{x_0+\varepsilon} f(x)\, v(x)\, dx > 0,$$

which contradicts the original assumption. An analogous argument rules out $f(x_0) < 0$ for some $a < x_0 < b$. *Q.E.D.*

Finite Elements Based on Weak Solutions

To characterize weak solutions, one imposes the appropriate integral criterion on the entire infinite-dimensional space of smooth test functions. Thus, an evident approximation strategy is to restrict the criterion to a suitable finite-dimensional subspace, thereby seeking an approximate weak solution that belongs to the subspace.

More precisely, concentrating on the self-adjoint framework discussed at the end of the preceding subsection, we restrict the weak formulation (10.66) of the linear system (10.65) to a finite-dimensional subspace $W \subset U$, and thus seek $w \in W$ such that

$$\langle\!\langle\, L[w]\,, L[\,v\,]\,\rangle\!\rangle = \langle\, f\,, v\,\rangle \qquad \text{for all} \qquad v \in W. \tag{10.69}$$

In this fashion, we characerize the *finite element approximation* to the weak solution u as the element $w \in W$ such that (10.69) holds for all $v \in W$.

To analyze this condition, as in (10.3), we now specify a basis $\varphi_1, \dots, \varphi_n$ of W, and thus can write both w and v as linear combinations thereof:

$$w = c_1\varphi_1 + \,\cdots\, + c_n\varphi_n, \qquad v = d_1\varphi_1 + \,\cdots\, + d_n\varphi_n.$$

Substituting these expressions into (10.69) produces the bilinear function

$$B(\mathbf{c}, \mathbf{d}) = \sum_{i,j=1}^{n} k_{ij} c_i d_j - \sum_{i=1}^{n} b_i d_i = \mathbf{c}^T K \mathbf{d} - \mathbf{b}^T \mathbf{d} = (K\mathbf{c} - \mathbf{b})^T \mathbf{d} = 0, \tag{10.70}$$

where

$$k_{ij} = \langle\!\langle L[\varphi_i], L[\varphi_j] \rangle\!\rangle, \qquad b_i = \langle f, \varphi_i \rangle, \qquad i, j = 1, \dots, n, \tag{10.71}$$

are the *same* as our earlier specifications (10.6, 7), and we used the fact that $K^T = K$ is a symmetric matrix to arrive at the final expression in (10.70). The condition that (10.69) hold for all $v \in W$ is equivalent to the requirement that (10.70) hold for all $\mathbf{d} = (d_1, d_2, \dots, d_n)^T \in \mathbb{R}^n$, which, in turn, implies that $\mathbf{c} = (c_1, c_2, \dots, c_n)^T$ must satisfy the linear system

$$K\mathbf{c} = \mathbf{b}.$$

But we immediately recognize that this is exactly the same as the finite element linear system (10.9)! We therefore conclude that *for a positive definite linear system constructed as above, the weak finite element approximation to the solution is the same as the minimizing finite element approximation.* In other words, it does not matter whether we characterize the solutions through the minimization principle or the weak reformulation; the resulting finite element approximations are exactly the same. There is thus no need to present any additional examples illustrating this construction.

In general, while the weak formulation is of much wider applicability, outside of boundary value problems with well-defined minimization principles, the rigorous underpinning that guarantees that the numerical solution is close to the actual solution is harder to establish and, in fact, not always valid. Indeed, one can find boundary value problems without analytic solutions that have spurious finite element numerical solutions, and, conversely, boundary value problems with solutions for which some finite element approximations do not exist because the resulting coefficient matrix is singular, [**113**, **126**].

Shock Waves as Weak Solutions

Finally, let us return to our earlier analysis, in Section 2.3, of shock waves, but now in the context of weak solutions. We begin by writing the nonlinear transport equation in the conservative form

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\tfrac{1}{2} u^2\right) = 0. \tag{10.72}$$

Since shock waves are discontinuous functions, they do not qualify as classical solutions. However, they can be rigorously characterized as weak solutions, a formulation that will, reassuringly, lead to the Rankine–Hugoniot Equal Area Rule for shock dynamics.

To construct a weak formulation of the nonlinear transport equation, we follow the general framework, and hence begin by multiplying the equation (10.72) by a smooth test function $v(t, x)$ and integrating over a domain $\Omega \subset \mathbb{R}^2$:

$$\iint_{\Omega} \left[\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\tfrac{1}{2} u^2\right) \right] v(t, x)\, dt\, dx = 0. \tag{10.73}$$

As a direct consequence of the two-dimensional version of the du Bois–Reymond Lemma, cf. Exercise 10.4.7, if $u(t, x) \in \mathrm{C}^1$ and condition (10.73) holds for all C^1 functions $v(t, x)$

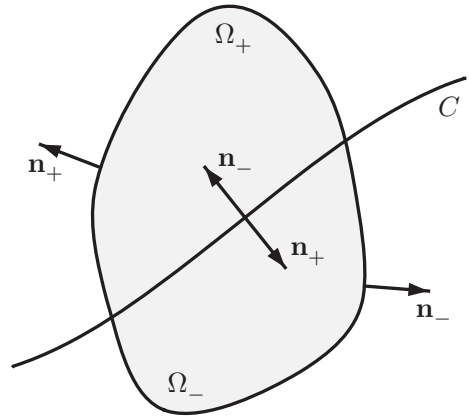


Figure 10.15. Integration domain for weak shock-wave solution.

with compact support contained in Ω, then $u(t,x)$ is necessarily a classical solution to the partial differential equation (10.72). The next step is to integrate by parts in order to remove the derivatives from u, and this is accomplished by appealing to *Green's formula* (6.82), which we rewrite in the form

$$\iint_\Omega \left(u_1 \frac{\partial v}{\partial t} + u_2 \frac{\partial v}{\partial x} \right) dt\, dx = \oint_{\partial\Omega} (\mathbf{u} \cdot \mathbf{n})\, v \; ds - \iint_\Omega \left(\frac{\partial u_1}{\partial t} + \frac{\partial u_2}{\partial x} \right) v \; dt\, dx, \quad (10.74)$$

where $\mathbf{u} = (\, u_1, u_2 \,)^T$. In our case, we identify the integral in (10.73) with the left-hand side of (10.74) by setting $u_1 = u$, $u_2 = \frac{1}{2} u^2$. Since v has compact support, the boundary integral vanishes, and thus we arrive at the weak formulation of the equation.

Definition 10.13. A function $u(t,x)$ is said to be a *weak solution* to the nonlinear transport equation (10.72) on $\Omega \subset \mathbb{R}^2$ if

$$\iint_\Omega \left(u \frac{\partial v}{\partial t} + \tfrac{1}{2} u^2 \frac{\partial v}{\partial x} \right) dt\, dx = 0 \quad (10.75)$$

for all C^1 functions $v(t,x)$ with compact support: $\operatorname{supp} v \subset \Omega$.

The key point is that, in the weak formulation (10.75), the derivatives are acting solely on $v(t,x)$, which we assume to be smooth, and not on our prospective solution $u(t,x)$, which now need not even be continuous for the integral to be well defined.

Let us derive the Rankine–Hugoniot shock condition (2.53) as a consequence of the weak formulation. Suppose $u(t,x)$ is a weak solution, defined on a domain $\Omega \subset \mathbb{R}^2$, that has a single jump discontinuity along a curve C parametrized by $x = \sigma(t)$ that separates Ω into two subdomains, say Ω_+ and Ω_-, such that its restriction to either subdomain, denoted by $u_+ = u \mid \Omega_+$ and $u_- = u \mid \Omega_-$, are each classical solutions on their respective domains, while the separating curve $C = \{\, x = \sigma(t) \,\}$ represents a shock-wave discontinuity. For specificity, we assume that Ω_+ lies above and Ω_- lies below C in the (t,x)–plane; see Figure 10.15. Let us investigate what the preceding weak formulation implies in this situation. We split the integral (10.75) into two parts, and then apply the integration by parts formula (10.74) to each individual double integral, keeping in mind that, when

restricted to Ω_+ or Ω_-, the integrand is sufficiently smooth to justify application of the formula:

$$\begin{aligned}
0 &= \iint_{\Omega} \left(u\,\frac{\partial v}{\partial t} + \tfrac{1}{2}\,u^2\,\frac{\partial v}{\partial x} \right) dt\,dx \\
&= \iint_{\Omega_+} \left(u_+\,\frac{\partial v}{\partial t} + \tfrac{1}{2}\,u_+^2\,\frac{\partial v}{\partial x} \right) dt\,dx \;+ \iint_{\Omega_-} \left(u_-\,\frac{\partial v}{\partial t} + \tfrac{1}{2}\,u_-^2\,\frac{\partial v}{\partial x} \right) dt\,dx \\
&= \oint_{\partial\Omega_+} (\widetilde{\mathbf{u}}_+ \cdot \mathbf{n}_+)\,v\,ds \;- \iint_{\Omega_+} \left[\frac{\partial u_+}{\partial t} + \frac{\partial}{\partial x}\left(\tfrac{1}{2}\,u_+^2 \right) \right] v\,dt\,dx + \\
&\qquad + \oint_{\partial\Omega_-} (\widetilde{\mathbf{u}}_- \cdot \mathbf{n}_-)\,v\,ds \;- \iint_{\Omega_-} \left[\frac{\partial u_-}{\partial t} + \frac{\partial}{\partial x}\left(\tfrac{1}{2}\,u_-^2 \right) \right] v\,dt\,dx \\
&= \int_C (\widetilde{\mathbf{u}}_+ \cdot \mathbf{n}_+ + \widetilde{\mathbf{u}}_- \cdot \mathbf{n}_-)\,v\,ds.
\end{aligned}$$

Here

$$\widetilde{\mathbf{u}}_+ = \begin{pmatrix} u_+ \\ \frac{1}{2}\,u_+^2 \end{pmatrix}, \qquad \widetilde{\mathbf{u}}_- = \begin{pmatrix} u_- \\ \frac{1}{2}\,u_-^2 \end{pmatrix},$$

while $\mathbf{n}_+, \mathbf{n}_-$ are the unit outwards normals on, respectively, $\partial\Omega_+$ and $\partial\Omega_-$. The final equality follows from the fact that the support of v is contained strictly inside Ω, and hence vanishes on those parts of the boundaries of Ω_+ and Ω_- that do not lie on the curve C. In particular, since C is the graph of $x = \sigma(t)$, the unit normals along it are, respectively,

$$\mathbf{n}_+ = \frac{1}{\sqrt{1 + \left(\dfrac{d\sigma}{dt}\right)^2}} \begin{pmatrix} \dfrac{d\sigma}{dt} \\ -1 \end{pmatrix}, \qquad \mathbf{n}_- = -\,\mathbf{n}_+ = \frac{1}{\sqrt{1 + \left(\dfrac{d\sigma}{dt}\right)^2}} \begin{pmatrix} -\dfrac{d\sigma}{dt} \\ 1 \end{pmatrix},$$

keeping in mind our convention that Ω_+ lies above and Ω_- lies below C, while

$$ds = \sqrt{1 + \left(\frac{d\sigma}{dt}\right)^2}\; dt.$$

Thus, the final line integral reduces to

$$\int_C \left[(u_- - u_+)\,\frac{d\sigma}{dt} - \tfrac{1}{2}\,(u_-^2 - u_+^2) \right] v\,dt = 0. \tag{10.76}$$

Since (10.76) vanishes for all C^1 functions $v(t,x)$ with compact support, the du Bois–Reymond Lemma 10.12 implies that

$$(u_- - u_+)\,\frac{d\sigma}{dt} = \tfrac{1}{2}\,(u_-^2 - u_+^2) \qquad \text{on} \qquad C,$$

thereby re-establishing the Rankine–Hugoniot shock condition (2.53). The upshot is that the shock-wave solutions produced in Section 2.3 are bona fide weak solutions.

Another computation shows that the rarefaction wave (2.54) also qualifies as a weak solution. However, so does the non-physical reverse shock solution discussed in Example 2.11. Thus, although the weak formulation recovers the Rankine–Hugoniot condition, it does not address the problem of causality, which must be additionally imposed to single

out a unique, physically meaningful weak solution. Further developments of these ideas can be found in more advanced monographs, e.g., [**107**, **122**].

Exercises

10.4.1. Write out semi-weak and fully weak formulations for the following boundary value problems: (*a*) $-u'' + 2u = x - x^2$, $u(0) = u(1) = 0$; (*b*) $e^x u'' + u = \cos x$, $u'(0) = u'(2) = 0$; (*c*) $x u'' + u' + x u = 0$, $u(1) = u(2) = 0$.

10.4.2. (*a*) Write down a weak formulation for the boundary value problem $-u'' + 3u = x$, $u(0) = u(1) = 0$. (*b*) Based on your weak formulation, construct a finite element approximation to the solution, using $n = 10$ nodes.

10.4.3. (*a*) Write down a weak formulation of the transport equation $u_t + 3u_x = 0$ on the real line. (*b*) Solve the initial value problem $u(0, x) = \begin{cases} 1 - |x|, & |x| \le 1, \\ 0, & \text{otherwise.} \end{cases}$

(*c*) Explain why the result of part (*b*) is not a classical solution to the wave equation. Is it a weak solution according to your formulation in part (*a*)?

10.4.4. (*a*) Write down a semi-weak formulation of the wave equation $u_{tt} = 4u_{xx}$ on the real line. (*b*) Solve the initial value problem $u(0, x) = \rho(x)$, $u_t(0, x) = 0$, where the initial displacement is a ramp function (6.25). (*c*) Explain why the result of part (*b*) is not a classical solution to the wave equation. Does it satisfy the semi-weak formulation of part (*a*)? Explain your answer.

♢ 10.4.5. (*a*) Starting with the nonlinear transport equation written in the alternative conservative form (2.56), find a corresponding weak formulation.
(*b*) Prove that your weak formulation produces the alternative entropy condition (2.58) for the motion of a shock discontinuity.

♢ 10.4.6. Prove that the du Bois–Reymond Lemma 10.12 remains valid even when $v(x) \in \mathrm{C}^\infty$ is required to be infinitely differentiable.

♢ 10.4.7. *The Two-dimensional du Bois–Reymond Lemma*: Let $\Omega \subset \mathbb{R}^2$ be a domain, and $f(t, x)$ a continuous function defined thereon. Prove that $\iint_\Omega f(t, x)\, v(t, x)\, dt\, dx = 0$ for every C^1 function $v(t, x)$ with compact support in Ω if and only if $f(t, x) \equiv 0$.

♠ 10.4.8. (*a*) Investigate the ability of finite elements to approximate a solution to the non-positive-definite boundary value problem $\Delta u + \lambda u = 0$, $0 < x < \pi$, $0 < y < \pi$, $u(x, 0) = 1$, $u(x, \pi) = u(0, y) = u(\pi, y) = 0$, when (*i*) $\lambda = 1$, (*ii*) $\lambda = 2$. Use separation of variables to find a series solution and use it to determine the accuracy of your finite element solution in part (*a*).