Numerical Analysis Lecture Notes

Peter J. Olver

14. Finite Elements

In this part, we introduce the powerful finite element method for finding numerical approximations to the solutions to boundary value problems involving both ordinary and partial differential equations can be solved by direct integration. The method relies on the characterization of the solution as the minimizer of a suitable quadratic functional. The innovative idea is to restrict the infinite-dimensional minimization principle characterizing the exact solution to a suitably chosen finite-dimensional subspace of the function space. When properly formulated, the solution to the resulting finite-dimensional minimization problem approximates the true minimizer. The finite-dimensional minimizer is found by solving the induced linear algebraic system, using either direct or iterative methods. We begin with one-dimensional boundary value problems involving ordinary differential equations, and, in the final section, show how to adapt the finite element analysis to partial differential equations, specifically the two-dimensional Laplace and Poisson equations.

14.1. Finite Elements for Ordinary Differential Equations.

The characterization of the solution to a linear boundary value problem via a quadratic minimization principle inspires a very powerful and widely used numerical solution scheme, known as the *finite element method*. In this final section, we give a brief introduction to the finite element method in the context of one-dimensional boundary value problems involving ordinary differential equations.

The underlying idea is strikingly simple. We are trying to find the solution to a boundary value problem by minimizing a quadratic functional $\mathcal{P}[u]$ on an infinite-dimensional vector space U. The solution $u_\star \in U$ to this minimization problem is found by solving a differential equation subject to specified boundary conditions. Minimizing the restriction of the the functional to a *finite-dimensional subspace* $W \subset U$ is a problem in linear algebra. Of course, restricting the functional $\mathcal{P}[u]$ to the subspace W will not, barring luck, lead to the exact minimizer. Nevertheless, if we choose W to be a sufficiently "large" subspace, the resulting minimizer $w_\star \in W$ may very well provide a reasonable approximation to the actual solution $u_\star \in U$. A rigorous justification of this process, under appropriate hypotheses, requires a full analysis of the finite element method, and we refer the interested reader to [**50**, **55**]. Here we shall concentrate on trying to understand how to apply the method in practice.

To be a bit more explicit, consider the minimization principle

$$\mathcal{P}[u] = \tfrac{1}{2} \,\|\, L[u] \,\|^2 - \langle\, f \,;\, u \,\rangle \tag{14.1}$$

for the linear system

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

representing our boundary value problem. The norm in (14.1) is typically based on some form of weighted inner product $\langle\!\langle\, v \,;\, \widetilde{v} \,\rangle\!\rangle$ on the space of strains $v = L[u] \in V$, while the inner product term $\langle\, f \,;\, u \,\rangle$ is typically (although not necessarily) unweighted on the space of displacements $u \in U$. The linear operator takes the self-adjoint form $K = L^* \circ L$, and must be positive definite — which requires $\ker L = \{0\}$. Without the positivity assumption, the boundary value problem has either no solutions, or infinitely many; in either event, the basic finite element method will not apply.

Rather than try to minimize $\mathcal{P}[u]$ on the entire function space U, we now seek to minimize it on a suitably chosen finite-dimensional subspace $W \subset U$. We begin by selecting a basis[†] $\varphi_1, \dots, \varphi_n$ of the subspace W. The general element of W is a (uniquely determined) linear combination

$$\varphi(x) = c_1\, \varphi_1(x) + \,\cdots\, + c_n\, \varphi_n(x) \tag{14.2}$$

of the basis functions. Our goal, then, is to determine the coefficients $c_1, \dots, c_n$ such that $\varphi(x)$ minimizes $\mathcal{P}[\varphi]$ among all such functions. Substituting (14.2) into (14.1) and expanding we find

$$\mathcal{P}[\varphi] = \frac{1}{2} \sum_{i,j=1}^{n} m_{ij}\, c_i\, c_j - \sum_{i=1}^{n} b_i\, c_i = \tfrac{1}{2}\, \mathbf{c}^T M\, \mathbf{c} - \mathbf{c}^T \mathbf{b}, \tag{14.3}$$

where

(a) $\mathbf{c} = (\, c_1, c_2, \dots, c_n \,)^T$ is the vector of unknown coefficients in (14.2),

(b) $M = (m_{ij})$ is the symmetric $n \times n$ matrix with entries

$$m_{ij} = \langle\!\langle\, L[\varphi_i] \,;\, L[\varphi_j] \,\rangle\!\rangle, \qquad i, j = 1, \dots, n, \tag{14.4}$$

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

$$b_i = \langle\, f \,;\, \varphi_i \,\rangle, \qquad i = 1, \dots, n. \tag{14.5}$$

Observe that, once we specify the basis functions φ_i, the coefficients m_{ij} and b_i are all known quantities. Therefore, we have reduced our original problem to a finite-dimensional problem of minimizing the quadratic function (14.3) over all possible vectors $\mathbf{c} \in \mathbb{R}^n$. The coefficient matrix M is, in fact, positive definite, since, by the preceding computation,

$$\mathbf{c}^T M\, \mathbf{c} = \sum_{i,j=1}^{n} m_{ij}\, c_i\, c_j = \|\, L[c_1\, \varphi_1(x) + \,\cdots\, + c_n\, \varphi_n] \,\|^2 = \|\, L[\varphi] \,\|^2 \,>\, 0 \tag{14.6}$$

[†] In this case, an orthonormal basis is not of any particular help.

as long as $L[\varphi] \neq 0$. Moreover, our positivity assumption implies that $L[\varphi] = 0$ if and only if $\varphi \equiv 0$, and hence (14.6) is indeed positive for all $\mathbf{c} \neq \mathbf{0}$. We can now invoke the original finite-dimensional minimization Theorem 12.12 to conclude that the unique minimizer to (14.3) is obtained by solving the associated linear system

$$M\mathbf{c} = \mathbf{b}. \tag{14.7}$$

Solving (14.7) relies on some form of Gaussian Elimination, or, alternatively, an iterative linear system solver, e.g., Gauss–Seidel or SOR.

This constitutes the basic abstract setting for the finite element method. The main issue, then, is how to effectively choose the finite-dimensional subspace W. Two candidates that might spring to mind are the space $\mathcal{P}^{(n)}$ of polynomials of degree $\leq n$, or the space $\mathcal{T}^{(n)}$ of trigonometric polynomials of degree $\leq n$, the focus of Fourier analysis. However, for a variety of reasons, neither is well suited to the finite element method. One criterion 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, the linear algebraic system (14.7) will typically be rather large, and so the coefficient matrix M should be as sparse as possible, i.e., have lots of zero entries. Otherwise, computing the solution will be too time-consuming to be of much practical value. Such considerations prove to be of absolutely crucial importance when applying the method to solve boundary value problems for partial differential equations in higher dimensions.

The really 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 $\mathcal{P}[u]$. The governing differential equation requires its solutions to have a certain degree of smoothness, whereas the associated minimization principle typically requires only half as many derivatives. Thus, for second order boundary value problems, e.g., Sturm–Liouville problems, $\mathcal{P}[u]$ only involves first order derivatives. It can be rigorously shown that the functional has the *same* minimizing solution, even if one allows (reasonable) functions that fail to have enough derivatives to satisfy the differential equation. Thus, one can try minimizing over subspaces containing fairly "rough" functions. Again, the justification of this method requires some deeper analysis, which lies beyond the scope of this introductory treatment.

For second order boundary value problems, a popular and effective choice of the finite-dimensional subspace is to use continuous, piecewise affine functions. Recall that a function is affine, $f(x) = a\,x + b$, if and only if its graph is a straight line. The function is *piecewise affine* if its graph consists of a finite number of straight line segments; a typical example is plotted in Figure 14.1. Continuity requires that the individual line 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 *mesh points*

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

The formulas simplify if one uses equally spaced mesh points, but this is not necessary for the method to apply. Let W denote the vector space consisting of all continuous, piecewise affine functions, with corners at the nodes, that satisfy the homogeneous boundary

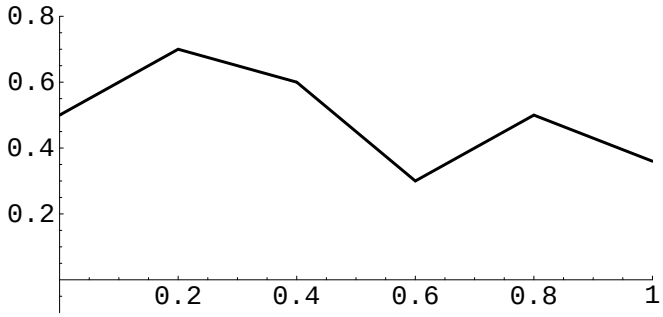

Figure 14.1. A Continuous Piecewise Affine Function.

conditions. To be specific, let us treat the case of Dirichlet (fixed) boundary conditions

$$\varphi(a) = \varphi(b) = 0. \tag{14.8}$$

Thus, on each subinterval

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

Continuity of $\varphi(x)$ requires

$$c_j = \varphi(x_j^+) = \varphi(x_j^-) = c_{j-1} + b_{j-1}\, h_{j-1}, \qquad j = 1, \dots, n-1, \tag{14.9}$$

where $h_{j-1} = x_j - x_{j-1}$ denotes the length of the j^{th} subinterval. The boundary conditions (14.8) require

$$\varphi(a) = c_0 = 0, \qquad \varphi(b) = c_{n-1} + h_{n-1}\, b_{n-1} = 0. \tag{14.10}$$

The function $\varphi(x)$ involves a total of $2n$ unspecified coefficients $c_0, \dots, c_{n-1}, b_0, \dots, b_{n-1}$. The continuity conditions (14.9) and the second boundary condition (14.10) uniquely determine the b_j. The first boundary condition specifies c_0, while the remaining $n-1$ coefficients $c_1 = \varphi(x_1), \dots, c_{n-1} = \varphi(x_{n-1})$ are arbitrary. We conclude that the finite element subspace W has dimension $n-1$, which is the number of interior mesh points.

Remark: Every function $\varphi(x)$ in our subspace has piecewise constant first derivative $w'(x)$. However, the jump discontinuities in $\varphi'(x)$ imply that its second derivative $\varphi''(x)$ has a delta function impulse at each mesh point, and is therefore far from being a solution to the differential equation. Nevertheless, the finite element minimizer $\varphi_\star(x)$ will, in practice, 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 that interpolate the basis data

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

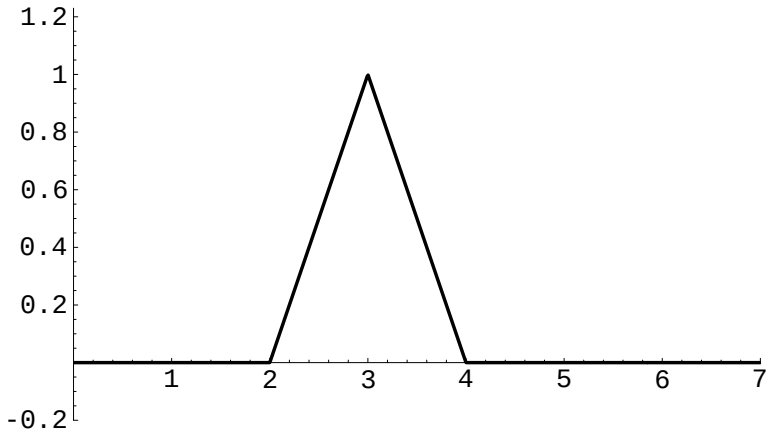

Figure 14.2. A Hat Function.

The graph of a typical hat function appears in Figure 14.2. The explicit formula is easily established:

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}}, & x_{j-1} \le x \le x_j, \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j}, & x_j \le x \le x_{j+1}, \\ 0, & x \le x_{j-1} \quad \text{or} \quad x \ge x_{j+1}, \end{cases} \qquad j = 1, \dots, n-1. \tag{14.11}$$

An advantage of using these basis elements is that the resulting coefficient matrix (14.4) turns out to be tridiagonal. Therefore, the tridiagonal Gaussian Elimination algorithm, [**42**], will rapidly produce the solution to the linear system (14.7). Since the accuracy of the finite element solution increases with the number of mesh points, this solution scheme allows us to easily compute very accurate numerical approximations.

Definition 14.1. 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 will belong to the support of $f(x)$, provided f is not zero there, or at least is not zero at nearby points.

Example 14.2. Consider the equilibrium equations

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

for a non-uniform bar subject to homogeneous Dirichlet boundary conditions. In order to formulate a finite element approximation scheme, we begin with the minimization principle based on the quadratic functional

$$\mathcal{P}[u] = \tfrac{1}{2}\,\|\,u'\,\|^2 - \langle\, f\,;u\,\rangle = \int_0^\ell \big[\,\tfrac{1}{2}\,c(x)\,u'(x)^2 - f(x)\,u(x)\,\big]\,dx.$$

We divide the interval $[\,0, \ell\,]$ into n equal subintervals, each of length $h = \ell/n$. The resulting uniform mesh has

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

The corresponding finite element basis hat functions are explicitly given by

$$\varphi_j(x) = \begin{cases} (x - x_{j-1})/h, & x_{j-1} \le x \le x_j, \\ (x_{j+1} - x)/h, & x_j \le x \le x_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \qquad\qquad j = 1, \dots, n-1. \tag{14.12}$$

The associated linear system (14.7) has coefficient matrix entries

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

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,

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

Therefore, the coefficient matrix has the tridiagonal form

$$M = \frac{1}{h^2} \begin{pmatrix} s_0 + s_1 & -s_1 & & & & \\ -s_1 & s_1 + s_2 & -s_2 & & & \\ & -s_2 & s_2 + s_3 & -s_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & -s_{n-3} & s_{n-3} + s_{n-2} & -s_{n-2} & \\ & & & -s_{n-2} & s_{n-2} + s_{n-1} \end{pmatrix}, \tag{14.13}$$

where

$$s_j = \int_{x_j}^{x_{j+1}} c(x)\, dx \tag{14.14}$$

is the total stiffness of the j^{th} subinterval. For example, in the homogeneous case $c(x) \equiv 1$, the coefficient matrix (14.13) reduces to the very special form

$$M = \frac{1}{h} \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}. \tag{14.15}$$

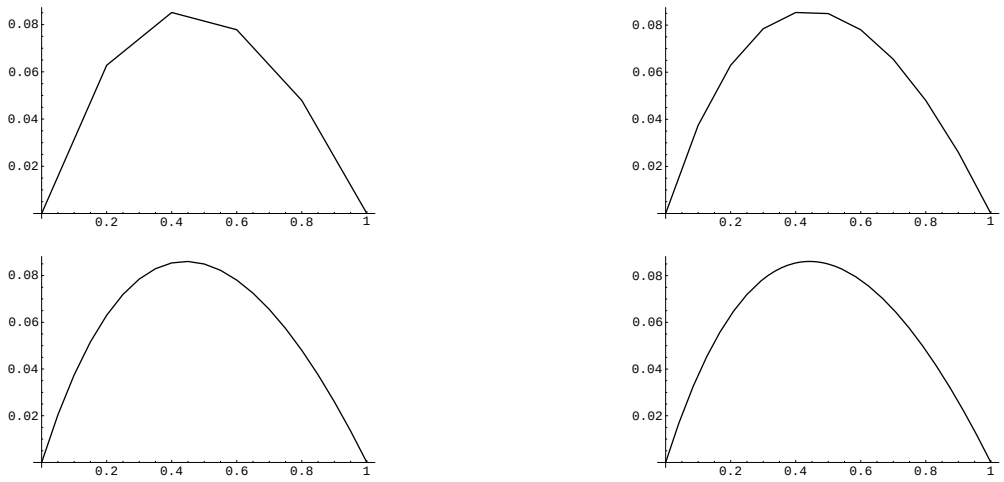

Figure 14.3. Finite Element Solution to (14.19).

The corresponding right hand side has entries

$$\begin{aligned} b_j = \langle\, f\,;\varphi_j\,\rangle &= \int_0^\ell f(x)\,\varphi_j(x)\,dx \\ &= \frac{1}{h}\left[\int_{x_{j-1}}^{x_j} (x - x_{j-1}) f(x)\,dx + \int_{x_j}^{x_{j+1}} (x_{j+1} - x) f(x)\,dx\right]. \end{aligned} \tag{14.16}$$

In this manner, we have assembled the basic ingredients for determining the finite element approximation to the solution.

In practice, we do not have to explicitly evaluate the integrals (14.14, 16), but may replace them by a suitably close numerical approximation. When $h \ll 1$ is small, then the integrals are taken over small intervals, and we can use the trapezoid rule†, [**7**], to approximate them:

$$s_j \;\approx\; \frac{h}{2}\big[\,c(x_j) + c(x_{j+1})\,\big], \qquad\qquad b_j \;\approx\; h\,f(x_j). \tag{14.17}$$

Remark: The j^{th} entry of the resulting finite element system $M\,\mathbf{c} = \mathbf{b}$ is, upon dividing by h, given by

$$-\,\frac{c_{j+1} - 2\,c_j + c_{j-1}}{h^2} \;=\; -\,\frac{u(x_{j+1}) - 2\,u(x_j) + u(x_{j-1})}{h^2} = -f(x_j). \tag{14.18}$$

The left hand side coincides with the standard finite difference approximation (11.6) to minus the second derivative $-u''(x_j)$ at the mesh point x_j. As a result, for this particular differential equation, the finite element and finite difference numerical solution methods happen to coincide.

† One might be tempted use more accurate numerical integration procedures, but the improvement in accuracy of the final answer is not very significant, particularly if the step size h is small.

Example 14.3. Consider the boundary value problem

$$-\frac{d}{dx}\,(x+1)\,\frac{du}{dx} = 1, \qquad u(0) = 0, \qquad u(1) = 0. \tag{14.19}$$

The explicit solution is easily found by direct integration:

$$u(x) = -x + \frac{\log(x+1)}{\log 2}\,. \tag{14.20}$$

It minimizes the associated quadratic functional

$$\mathcal{P}[\,u\,] = \int_0^\ell \left[\, \tfrac{1}{2}\,(x+1)\,u'(x)^2 - u(x)\,\right] dx \tag{14.21}$$

over all possible functions $u \in \mathrm{C}^1$ that satisfy the given boundary conditions. The finite element system (14.7) has coefficient matrix given by (14.13) and right hand side (14.16), where

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

The resulting solution is plotted in Figure 14.3. The first three graphs contain, respectively, 5, 10, 20 points in the mesh, so that $h = .2, .1, .05$, while the last plots the exact solution (14.20). Even when computed on rather coarse meshes, the finite element approximation is quite respectable.

Example 14.4. Consider the Sturm–Liouville boundary value problem

$$-u'' + (x+1)u = x\,e^x, \qquad u(0) = 0, \qquad u(1) = 0. \tag{14.22}$$

The solution minimizes the quadratic functional

$$\mathcal{P}[\,u\,] = \int_0^1 \left[\, \tfrac{1}{2}\,u'(x)^2 + \tfrac{1}{2}\,(x+1)\,u(x)^2 - e^x\,u(x)\,\right] dx, \tag{14.23}$$

over all functions $u(x)$ that satisfy the boundary conditions. We lay out a uniform mesh of step size $h = 1/n$. The corresponding finite element basis hat functions as in (14.12). The matrix entries are given by[†]

$$m_{ij} = \int_0^1 \left[\, \varphi_i'(x)\,\varphi_j'(x) + (x+1)\,\varphi_i(x)\,\varphi_j(x)\,\right] dx \;\approx \begin{cases} \dfrac{2}{h} + \dfrac{2\,h}{3}\,(x_i+1), & i = j, \\ -\dfrac{1}{h} + \dfrac{h}{6}\,(x_i+1), & |\,i-j\,| = 1, \\ 0, & \text{otherwise}, \end{cases}$$

[†] The integration is made easier by noting that the integrand is zero except on a small subinterval. Since the function $x+1$ (but not φ_i or φ_j) does not vary significantly on this subinterval, it can be approximated by its value $1 + x_i$ at a mesh point. A similar simplification is used in the ensuing integral for b_i.

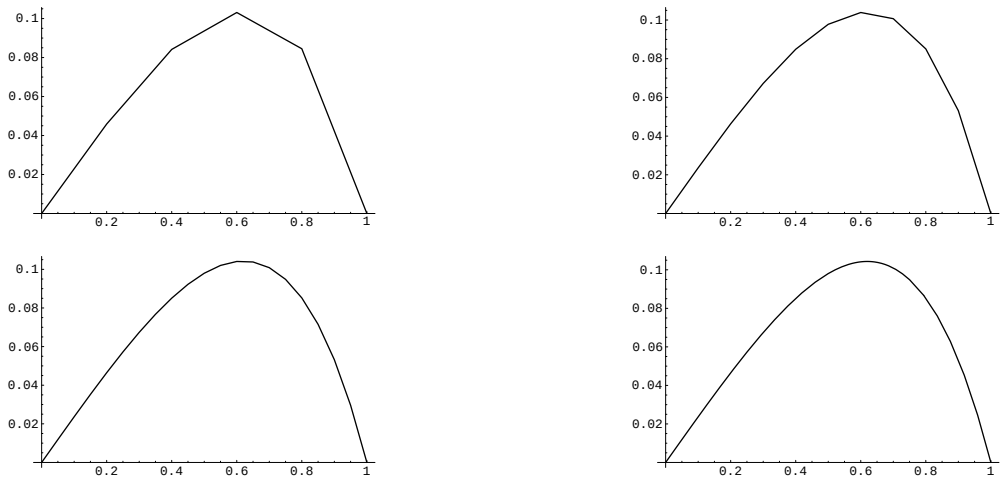

Figure 14.4. Finite Element Solution to (14.22).

while

$$b_i = \langle\, x\, e^x \,;\, \varphi_i \,\rangle = \int_0^1 x\, e^x\, \varphi_i(x)\, dx \;\approx\; x_i\, e^{x_i}\, h.$$

The resulting solution is plotted in Figure 14.4. As in the previous figure, the first three graphs contain, respectively, 5, 10, 20 points in the mesh, while the last plots the exact solution, which can be expressed in terms of Airy functions, cf. [**40**].

So far, we have only treated homogeneous boundary conditions. An inhomogeneous boundary value problem does not immediately fit into our framework since the set of functions satisfying the boundary conditions does *not* form a vector space. One way to get around this problem is to replace $u(x)$ by $\widetilde{u}(x) = u(x) - h(x)$, where $h(x)$ is any convenient function that satisfies the boundary conditions. For example, for the inhomogeneous Dirichlet conditions

$$u(a) = \alpha, \qquad\qquad u(b) = \beta,$$

we can subtract off the affine function

$$h(x) = \frac{(\beta - \alpha)x + \alpha\, b - \beta\, a}{b - a}.$$

Another option is to choose an appropriate combination of elements at the endpoints:

$$h(x) = \alpha\, \varphi_0(x) + \beta\, \varphi_n(x).$$

Linearity implies that the difference $\widetilde{u}(x) = u(x) - h(x)$ satisfies the amended differential equation

$$K[\widetilde{u}] = \widetilde{f}, \qquad\qquad \text{where} \qquad\qquad \widetilde{f} = f - K[h],$$

now supplemented by homogeneous boundary conditions. The modified boundary value problem can then be solved by the standard finite element method. Further details are left as a project for the motivated student.

Finally, one can employ other functions beyond the piecewise affine hat functions (14.11) to span finite element subspace. Another popular choice, which is essential for

higher order boundary value problems such as beams, is to use splines. Thus, once we have chosen our mesh points, we can let $\varphi_j(x)$ be the basis B–splines discussed in [**42**]. Since $\varphi_j(x) = 0$ for $x \le x_{j-2}$ or $x \ge x_{j+2}$, the resulting coefficient matrix (14.4) is *pentadiagonal*, which means $m_{ij} = 0$ whenever $|\, i - j \,| > 2$. Pentadiagonal matrices are not quite as pleasant as their tridiagonal cousins, but are still rather sparse. Positive definiteness of M implies that an iterative solution technique, e.g., SOR, can effectively and rapidly solve the linear system, and thereby produce the finite element spline approximation to the boundary value problem.

14.2. Finite Elements in Two Dimensions.

Finite element methods are also effectively employed to solving boundary value problems for elliptic partial differential equations. In this section, we concentrate on applying these ideas to the two-dimensional Poisson equation. For specificity, we concentrate on the homogeneous Dirichlet boundary value problem.

Theorem 14.5. *The function $u(x, y)$ that minimizes the* Dirichlet integral

$$\tfrac{1}{2} \,\|\, \nabla u \,\|^2 - \langle\, u \,;\, f \,\rangle = \iint_\Omega \left(\tfrac{1}{2}\, u_x^2 + \tfrac{1}{2}\, u_y^2 - f\, u \right) dx\, dy \tag{14.24}$$

among all C^1 functions that satisfy the prescribed homogeneous Dirichlet boundary conditions is the solution to the boundary value problem

$$-\Delta u = f \quad \text{in} \quad \Omega \qquad\qquad u = 0 \quad \text{on} \quad \partial\Omega. \tag{14.25}$$

In the finite element approximation, we restrict the Dirichlet functional to a suitably chosen finite-dimensional subspace. As in the one-dimensional situation, the most convenient finite-dimensional subspaces consist of functions that may lack the requisite degree of smoothness that qualifies them as possible solutions to the partial differential equation. Nevertheless, they do provide good approximations to the actual solution. An important practical consideration, impacting the speed of the calculation, is to employ functions with small support. The resulting finite element matrix will then be sparse and the solution to the linear system can be relatively rapidly calculate, usually by application of an iterative numerical scheme such as the Gauss–Seidel or SOR methods discussed in Section 7.4.

Finite Elements and Triangulation

For one-dimensional boundary value problems, the finite element construction rests on the introduction of a mesh $a = x_0 < x_1 < \cdots < x_n = b$ on the interval of definition. The mesh nodes x_k break the interval into a collection of small subintervals. In two-dimensional problems, a *mesh* consists of a finite number of points $\mathbf{x}_k = (x_k, y_k)$, $k = 1, \dots, m$, known as *nodes*, usually lying inside the domain $\Omega \subset \mathbb{R}^2$. As such, there is considerable freedom in the choice of mesh nodes, and completely uniform spacing is often not possible. We regard the nodes as forming the vertices of a *triangulation* of the domain Ω, consisting of a finite number of small triangles, which we denote by $T_1, \dots, T_N$. The nodes are split into two categories — *interior nodes* and *boundary nodes*, the latter lying on or close to the boundary of the domain. A curved boundary is approximated by the polygon through the

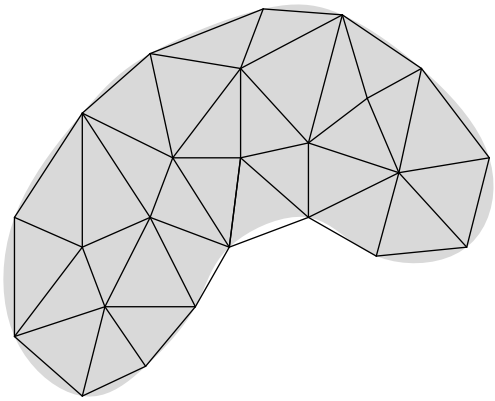

Figure 14.5. Triangulation of a Planar Domain.

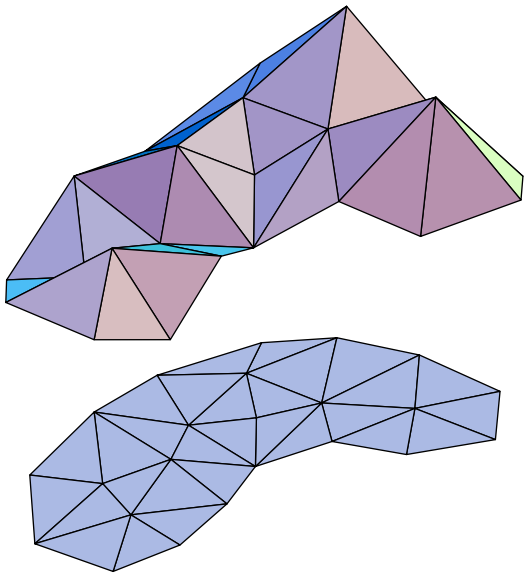

Figure 14.6. Piecewise Affine Function.

boundary nodes formed by the sides of the triangles lying on the edge of the domain; see Figure 14.5 for a typical example. Thus, in computer implementations of the finite element method, the first module is a routine that will automatically triangulate a specified domain in some reasonable manner; see below for details on what "reasonable" entails.

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

$$w(x, y) = \alpha^\nu + \beta^\nu x + \gamma^\nu y, \qquad \text{for} \qquad (x, y) \in T_\nu. \tag{14.26}$$

Continuity of w requires that its values on a common edge between two triangles must

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

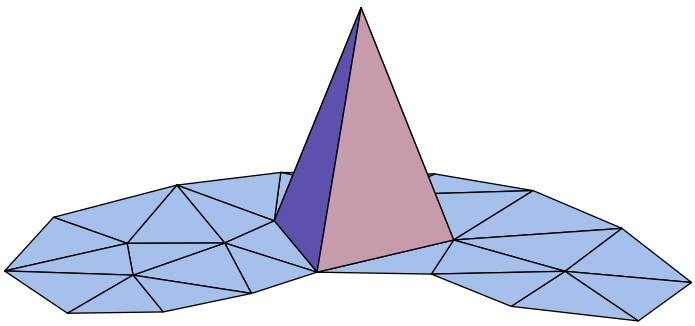

Figure 14.7. Finite Element Pyramid Function.

agree, and this will impose certain compatibility conditions on the coefficients $\alpha^\mu, \beta^\mu, \gamma^\mu$ and $\alpha^\nu, \beta^\nu, \gamma^\nu$ associated with adjacent pairs of triangles T_μ, T_ν. The graph of $z = w(x, y)$ forms a connected polyhedral surface whose triangular faces lie above the triangles in the domain; see Figure 14.6 for an illustration.

The next step is to choose a basis of the subspace of piecewise affine functions for the given triangulation. As in the one-dimensional version, the most convenient basis consists of *pyramid functions* $\varphi_k(x, y)$ which assume the value 1 at a single node $\mathbf{x}_k$, and are zero at all the other nodes; thus

$$\varphi_k(x_i, y_i) = \begin{cases} 1, & i = k, \\ 0, & i \neq k. \end{cases} \tag{14.27}$$

Note that φ_k will be nonzero only on those triangles which have the node $\mathbf{x}_k$ as one of their vertices, and hence the graph of φ_k looks like a pyramid of unit height sitting on a flat plane, as illustrated in Figure 14.7.

The pyramid functions $\varphi_k(x, y)$ corresponding to the *interior nodes* $\mathbf{x}_k$ automatically satisfy the homogeneous Dirichlet boundary conditions on the boundary of the domain — or, more correctly, on the polygonal boundary of the triangulated domain, which is supposed to be a good approximation to the curved boundary of the original domain Ω. Thus, the finite-dimensional finite element subspace W is the span of the interior node pyramid functions, and so general element $w \in W$ is a linear combination thereof:

$$w(x, y) = \sum_{k=1}^{n} c_k \, \varphi_k(x, y), \tag{14.28}$$

where the sum ranges over the n interior nodes of the triangulation. Owing to the original specification (14.27) of the pyramid functions, the coefficients

$$c_k = w(x_k, y_k) \approx u(x_k, y_k), \qquad k = 1, \dots, n, \tag{14.29}$$

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 $c_1 = \cdots = c_n = 0$.

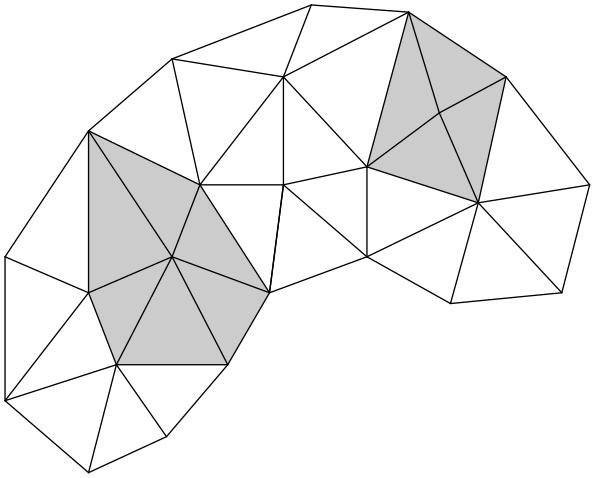

Figure 14.8. Vertex Polygons.

Thus, the interior node pyramid functions $\varphi_1, \dots \varphi_n$ form a basis for finite element subspace W, which therefore has dimension equal to n, the number of interior nodes.

Determining the explicit formulae for the finite element basis functions is not difficult. On one of the triangles T_ν that has $\mathbf{x}_k$ as a vertex, $\varphi_k(x, y)$ will be the unique affine function (14.26) that takes the value 1 at the vertex $\mathbf{x}_k$ and 0 at its other two vertices $\mathbf{x}_l$ and $\mathbf{x}_m$. Thus, we are in need of a formula for an affine function or *element*

$$\omega_k^\nu(x, y) = \alpha_k^\nu + \beta_k^\nu\, x + \gamma_k^\nu\, y, \qquad\qquad (x, y) \in T_\nu, \tag{14.30}$$

that takes the prescribed values

$$\omega_k^\nu(x_i, y_i) = \omega_k^\nu(x_j, y_j) = 0, \qquad\qquad \omega_k^\nu(x_k, y_k) = 1,$$

at three distinct points. These three conditions lead to the linear system

$$\begin{aligned} \omega_k^\nu(x_i, y_i) &= \alpha_k^\nu + \beta_k^\nu\, x_i + \gamma_k^\nu\, y_i \;\; = 0, \\ \omega_k^\nu(x_j, y_j) &= \alpha_k^\nu + \beta_k^\nu\, x_j + \gamma_k^\nu\, y_j \;\; = 0, \\ \omega_k^\nu(x_k, y_k) &= \alpha_k^\nu + \beta_k^\nu\, x_k + \gamma_k^\nu\, y_k = 1. \end{aligned} \tag{14.31}$$

The solution produces the explicit formulae

$$\alpha_k^\nu = \frac{x_i\, y_j - x_j\, y_i}{\Delta_\nu}\,, \qquad\qquad \beta_k^\nu = \frac{y_i - y_j}{\Delta_\nu}\,, \qquad\qquad \gamma_k^\nu = \frac{x_j - x_i}{\Delta_\nu}\,, \tag{14.32}$$

for the coefficients; the denominator

$$\Delta_\nu = \det \begin{pmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{pmatrix} = \pm 2 \text{ area } T_\nu \tag{14.33}$$

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

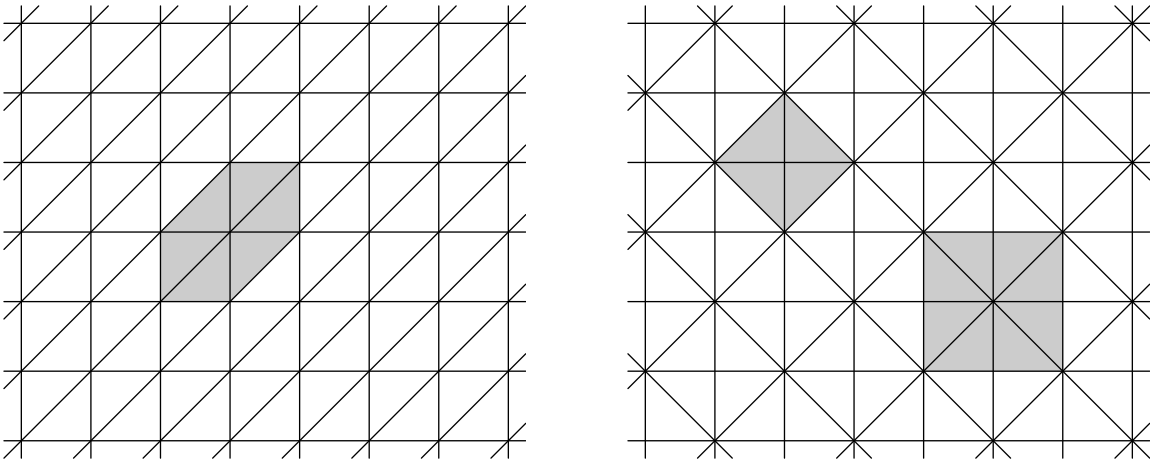

Figure 14.9. Square Mesh Triangulations.

Example 14.6. Consider an isoceles right triangle T with vertices

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

Using (14.32–33) (or solving the linear systems (14.31) directly), we immediately produce the three affine elements

$$\omega_1(x,y) = 1 - x - y, \qquad \omega_2(x,y) = x, \qquad \omega_3(x,y) = y. \tag{14.34}$$

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

The finite element pyramid function is then obtained by piecing together the individual affine elements, whence

$$\varphi_k(x,y) = \begin{cases} \omega_k^{\nu}(x,y), & \text{if } (x,y) \in T_\nu \text{ which has } \mathbf{x}_k \text{ as a vertex,} \\ 0, & \text{otherwise.} \end{cases} \tag{14.35}$$

Continuity of $\varphi_k(x,y)$ is assured since the constituent affine elements have the same values at common vertices. The support of the pyramid function (14.35) is the polygon

$$\operatorname{supp} \varphi_k = P_k = \bigcup_{\nu} T_\nu \tag{14.36}$$

consisting of all the triangles T_ν that have the node $\mathbf{x}_k$ as a vertex. In other words, $\varphi_k(x,y) = 0$ whenever $(x,y) \notin P_k$. We will call P_k the k^{th} *vertex polygon*. The node $\mathbf{x}_k$ lies on the interior of its vertex polygon P_k, while the vertices of P_k are all those that are connected to $\mathbf{x}_k$ by a single edge of the triangulation. In Figure 14.8 the shaded regions indicate two of the vertex polygons for the triangulation in Figure 14.5.

Example 14.7. The simplest, and most common triangulations are based on regular meshes. Suppose that the nodes lie on a square grid, and so are of the form $\mathbf{x}_{i,j} = (ih + a, jh + b)$ where $h > 0$ 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 14.9, then the vertex polygons all have the same shape, consisting of 6 triangles

of total area $3h^2$ — the shaded region. On the other hand, if we choose an alternating, perhaps more æsthetically pleasing triangulation as in the second picture, then there are two types of vertex polygons. The first, consisting of four triangles, has area $2h^2$, while the second, containing 8 triangles, has twice the area, $4h^2$. In practice, there are good reasons to prefer the former triangulation.

In general, in order to ensure convergence of the finite element solution to the true minimizer, one should choose a triangulation with the following properties:

(a) The triangles are not too long and skinny. In other words, their sides should have comparable lengths. In particular, obtuse triangles should be avoided.

(b) The areas of nearby triangles T_ν should not vary too much.

(c) The areas of nearby vertex polygons P_k should also not vary too much.

For adaptive or variable meshes, one might very well have wide variations in area over the entire grid, with small triangles in regions of rapid change in the solution, and large ones in less interesting regions. But, overall, the sizes of the triangles and vertex polygons should not dramatically vary as one moves across the domain.

The Finite Element Equations

We now seek to approximate the solution to the homogeneous Dirichlet boundary value problem by restricting the Dirichlet functional to the selected finite element subspace W. Substituting the formula (14.28) for a general element of W into the quadratic Dirichlet functional (14.24) and expanding, we find

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

Here, $K = (k_{ij})$ is the symmetric $n \times n$ matrix, while $\mathbf{b} = (\,b_1, b_2, \ldots, b_n\,)^T$ is the vector that have the respective entries

$$\begin{aligned}k_{ij} &= \langle\,\nabla\varphi_i\,;\nabla\varphi_j\,\rangle = \iint_\Omega \nabla\varphi_i\cdot\nabla\varphi_j\,dx\,dy,\\ b_i &= \langle\,f\,;\varphi_i\,\rangle = \iint_\Omega f\,\varphi_i\,dx\,dy.\end{aligned} \tag{14.37}$$

Thus, to determine the finite element approximation, we need to minimize the quadratic function

$$P(\mathbf{c}) = \tfrac{1}{2}\,\mathbf{c}^T K\,\mathbf{c} - \mathbf{b}^T\mathbf{c} \tag{14.38}$$

over all possible choices of coefficients $\mathbf{c} = (\,c_1, c_2, \ldots, c_n\,)^T \in \mathbb{R}^n$, i.e., over all possible function values at the interior nodes. Restricting to the finite element subspace has reduced us to a standard finite-dimensional quadratic minimization problem. First, the coefficient matrix $K > 0$ is positive definite due to the positive definiteness of the original functional;

the proof in Section 14.1 is easily adapted to the present situation. Theorem 12.12 tells us that the minimizer is obtained by solving the associated linear system

$$K\,\mathbf{c} = \mathbf{b}. \tag{14.39}$$

The solution to (14.39) can be effected by either Gaussian elimination or an iterative technique.

To find explicit formulae for the matrix coefficients k_{ij} in (14.37), we begin by noting that the gradient of the affine element (14.30) is equal to

$$\nabla \omega_k^\nu(x,y) = \mathbf{a}_k^\nu = \begin{pmatrix} \beta_k^\nu \\ \gamma_k^\nu \end{pmatrix} = \frac{1}{\Delta_\nu} \begin{pmatrix} y_i - y_j \\ x_j - x_i \end{pmatrix}, \qquad (x,y) \in T_\nu, \tag{14.40}$$

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

$$\nabla \varphi_k(x,y) = \begin{cases} \nabla \omega_k^\nu = \mathbf{a}_k^\nu, & \text{if } (x,y) \in T_\nu \text{ which has } \mathbf{x}_k \text{ as a vertex,} \\ \mathbf{0}, & \text{otherwise,} \end{cases} \tag{14.41}$$

reduces to a piecewise constant function on the triangulation. Actually, (14.41) is not quite correct since if (x,y) lies on the boundary of a triangle T_ν, then the gradient does not exist. However, this technicality will not cause any difficulty in evaluating the ensuing integral.

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

$$k_{ij} \;\approx\; \sum_\nu \iint_{T_\nu} \nabla \varphi_i \cdot \nabla \varphi_j \, dx\, dy \;\equiv\; \sum_\nu k_{ij}^\nu. \tag{14.42}$$

Now, according to (14.41), 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, 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 (14.41),

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

Let T_ν have vertices $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$. Then, by (14.40, 41, 33),

$$\begin{aligned} k_{ij}^\nu &= \frac{1}{2}\, \frac{(y_j - y_k)(y_k - y_i) + (x_k - x_j)(x_i - x_k)}{(\Delta_\nu)^2}\, |\, \Delta_\nu \,| = -\,\frac{(\mathbf{x}_i - \mathbf{x}_k)\cdot(\mathbf{x}_j - \mathbf{x}_k)}{2\,|\,\Delta_\nu\,|}\,, \quad i \neq j, \\ k_{ii}^\nu &= \frac{1}{2}\, \frac{(y_j - y_k)^2 + (x_k - x_j)^2}{(\Delta_\nu)^2}\, |\, \Delta_\nu \,| = \frac{\|\, \mathbf{x}_j - \mathbf{x}_k \,\|^2}{2\,|\,\Delta_\nu\,|}\,. \end{aligned} \tag{14.43}$$

In this manner, each triangle T_ν specifies a collection of 6 different coefficients, $k_{ij}^\nu = k_{ji}^\nu$, indexed by its vertices, and known as the *elemental stiffnesses* of T_ν. Interestingly, the

Figure 14.10. Right and Equilateral Triangles.

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, if $\theta_i^\nu, \theta_j^\nu, \theta_k^\nu$ denote the angles in T_ν at the respective vertices $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$, then, according to Exercise ■,

$$k_{ii}^\nu = \tfrac{1}{2}\big(\cot\theta_k^\nu + \cot\theta_j^\nu\big), \qquad \text{while} \qquad k_{ij}^\nu = k_{ji}^\nu = -\tfrac{1}{2}\cot\theta_k^\nu, \quad i \neq j. \tag{14.44}$$

Example 14.8. The right triangle with vertices $\mathbf{x}_1 = (0,0)$, $\mathbf{x}_2 = (1,0)$, $\mathbf{x}_3 = (0,1)$ has elemental stiffnesses

$$k_{11} = 1, \qquad k_{22} = k_{33} = \tfrac{1}{2}, \qquad k_{12} = k_{21} = k_{13} = k_{31} = -\tfrac{1}{2}, \qquad k_{23} = k_{32} = 0. \tag{14.45}$$

The same holds for any other isoceles right triangle, as long as we chose the first vertex to be at the right angle. Similarly, an equilateral triangle has all 60° angles, and so its elemental stiffnesses are

$$\begin{aligned} &k_{11} = k_{22} = k_{33} = \tfrac{1}{\sqrt{3}} \approx .577350, \\ &k_{12} = k_{21} = k_{13} = k_{31} = k_{23} = k_{32} = -\tfrac{1}{2\sqrt{3}} \approx -.288675. \end{aligned} \tag{14.46}$$

Assembling the Elements

The elemental stiffnesses of each triangle will contribute, through the summation (14.42), to the finite element coefficient matrix K. We begin by constructing a larger matrix 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 K^* are labeled by the nodes $\mathbf{x}_i$. 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) 9 nonzero entries. The resulting $m \times m$ matrices are all summed together over all the triangles,

$$K^* = \sum_{\nu=1}^{N} K_\nu, \tag{14.47}$$

to produce the full finite element matrix, in accordance with (14.42).

The full finite element matrix K^* is too large, since its rows and columns include all the nodes, whereas the finite element matrix K appearing in (14.39) only refers to the n interior nodes. The *reduced* $n \times n$ *finite element matrix* K is simply obtained from K^* by

Figure 14.11. The Oval Plate.

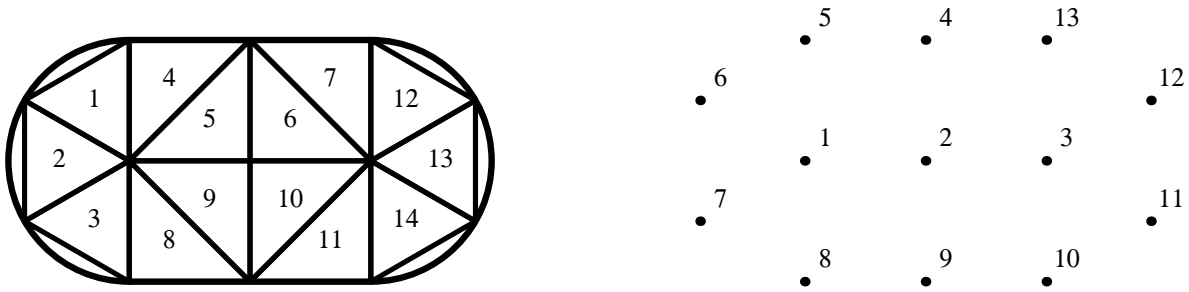

Figure 14.12. A Coarse Triangulation of the Oval Plate.

deleting all rows and columns indexed by boundary nodes, retaining only the elements k_{ij} when both $\mathbf{x}_i$ and $\mathbf{x}_j$ are interior nodes. For the homogeneous boundary value problem, this is all we require. As we shall see, inhomogeneous boundary conditions are most easily handled by retaining (part of) the full matrix K^*.

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

Example 14.9. 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 semicircular disks glued onto its shorter ends, as sketched in Figure 14.11. 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 Poisson equation with Dirichlet boundary conditions, for the equilibrium temperature $u(x, y)$.

Let us describe how to set up the finite element approximation to such a boundary value problem. We begin with a very coarse triangulation of the plate, which will not give particularly accurate results, but does serve to illustrate how to go about assembling the finite element matrix. We divide the rectangular part of the plate into 8 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 14.12. 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, going counterclockwise around the boundary starting at the top. The full finite element matrix 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 (14.39) consists of the upper left 3×3 submatrix of K^* corresponding to the three interior nodes.

Each triangle T_ν will contribute the summand K_ν whose values are its elemental stiffnesses, as indexed by its vertices. For example, the first triangle T_1 is equilateral, and so has elemental stiffnesses (14.46). Its vertices are labeled 1, 5, and 6, and therefore we place the stiffnesses (14.46) in the rows and columns numbered $1, 5, 6$ to form the summand

$$K_1 = \begin{pmatrix} .577350 & 0 & 0 & 0 & -.288675 & -.288675 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ -.288675 & 0 & 0 & 0 & .577350 & -.288675 & 0 & 0 & \dots \\ -.288675 & 0 & 0 & 0 & -.288675 & .577350 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

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 (14.46), but now its vertices are $1, 6, 7$, and so it will contribute

$$K_2 = \begin{pmatrix} .577350 & 0 & 0 & 0 & 0 & -.288675 & -.288675 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ -.288675 & 0 & 0 & 0 & 0 & .577350 & -.2886750 & 0 & \dots \\ -.288675 & 0 & 0 & 0 & 0 & -.288675 & .5773500 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Similarly for K_3, with vertices $1, 7, 8$. On the other hand, triangle T_4 is an isoceles right triangle, and so has elemental stiffnesses (14.45). 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 the sum

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

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

$$\begin{pmatrix} -.7887 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -.7887 & -.5774 & -.5774 & -.7887 \\ 0 & 0 & 0 & 0 & 0 & -.5 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -.2887 & 0 & 0 & 0 & 0 & 0 \\ 1.577 & -.5 & 0 & 0 & 0 & 0 \\ -.5 & 2 & -.5 & 0 & 0 & 0 \\ 0 & -.5 & 1.577 & -.2887 & 0 & 0 \\ 0 & 0 & -.2887 & 1.155 & -.2887 & 0 \\ 0 & 0 & 0 & -.2887 & 1.155 & -.2887 \\ 0 & 0 & 0 & 0 & -.2887 & 1.577 \end{pmatrix}. \tag{14.48}$$

Since only nodes $1, 2, 3$ are interior nodes, the reduced finite element matrix only uses the upper left 3×3 block of K^*, so

$$K = \begin{pmatrix} 3.732 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 3.732 \end{pmatrix}. \tag{14.49}$$

It is not difficult to directly construct K, bypassing 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 is when we use a uniform square mesh, and so triangulate the domain into isoceles 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 14.13 shows the nodes in a square grid with intermesh spacing $h = .2$. While a bit crude in its approximation

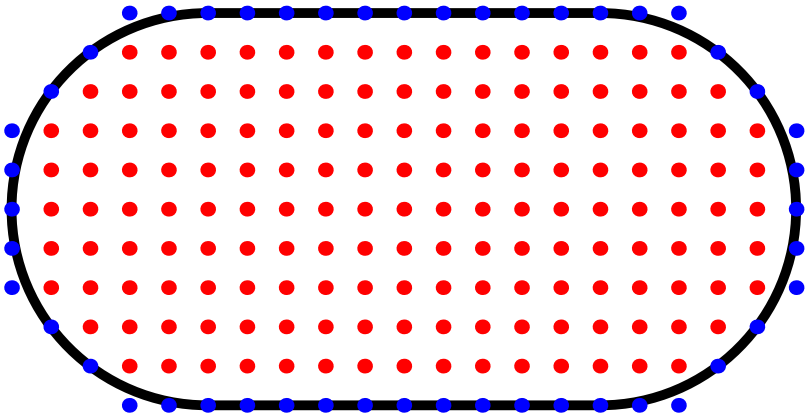

Figure 14.13. A Square Mesh for the Oval Plate.

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 isoceles right triangles, with elemental stiffnesses (14.45). Summing the corresponding matrices K_ν over all the triangles, as in (14.47), the rows and columns of K^* corresponding to the interior nodes are seen to all have the same 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 either -1 when node i is adjacent to node j on the grid, and is 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 only plays a role 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 isoceles triangles, and the corresponding entries in (14.48) are $k_{22} = 4$, and $k_{2j} = -1$ for the four adjacent nodes $j = 1, 3, 4, 9$.

Remark: Interestingly, the coefficient matrix arising 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 Exercise ■. The finite element approach has the advantage of applying to much more general triangulations.

In general, while the finite element matrix K for a two-dimensional boundary value problem is not as nice as the tridiagonal matrices we obtained in our one-dimensional problems, it is still very sparse and, on regular grids, highly structured. This makes solution of the resulting linear system particularly amenable to an iterative matrix solver such as Gauss–Seidel, Jacobi, or, for even faster convergence, successive over-relaxation (SOR).

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

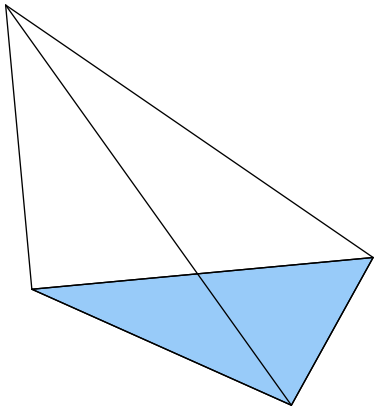

Figure 14.14. Finite Element Tetrahedron.

hand side of the fundamental linear equation (14.39). According to (14.37), 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

$$b_i = \iint_\Omega f\,\varphi_i\,dx\,dy \approx \sum_\nu \iint_{T_\nu} f\,\omega_i^\nu\,dx\,dy \equiv \sum_\nu b_i^\nu. \tag{14.50}$$

Typically, the exact computation of the various triangular integrals is not convenient, and so we resort to a numerical approximation. Since we are assuming that the individual triangles are small, we can adopt 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 $(x,y) \in T_\nu$ by a constant. The integral (14.50) is then approximated by

$$b_i^\nu = \iint_{T_\nu} f\,\omega_i^\nu\,dx\,dy \approx c_i^\nu \iint_{T_\nu} \omega_i^\nu(x,y)\,dx\,dy = \tfrac{1}{3}\,c_i^\nu \text{ area } T_\nu = \tfrac{1}{6}\,c_i^\nu\,|\,\Delta_\nu\,|. \tag{14.51}$$

The formula for the integral of the affine element $\omega_i^\nu(x,y)$ follows from solid geometry. Indeed, it equals the volume under its graph, a tetrahedron of height 1 and base T_ν, as illustrated in Figure 14.14.

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 (14.50) becomes

$$b_i \approx \sum_\nu \tfrac{1}{3}\,f(x_i,y_i) \text{ area } T_\nu = \tfrac{1}{3}\,f(x_i,y_i) \text{ area } P_i, \tag{14.52}$$

where P_i is the vertex polygon (14.36) corresponding to the node $\mathbf{x}_i$. In particular, for the square mesh with the uniform choice of triangles, as in Example 14.7,

$$\text{area } P_i = 3\,h^2 \qquad \text{for all } i, \text{ and so} \qquad b_i \approx f(x_i,y_i)\,h^2 \tag{14.53}$$

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

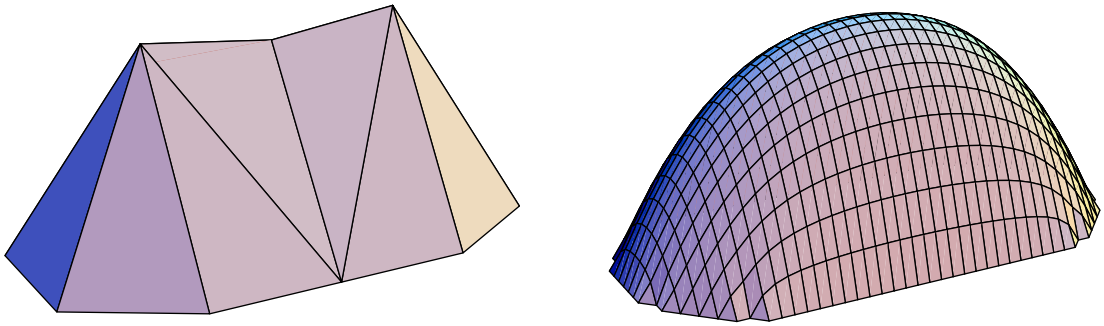

Figure 14.15. Finite Element Solutions to Poisson's Equation for an Oval Plate.

alternating version would give unequal values for the b_i over adjacent nodes, and this would introduce unnecessary errors into the final approximation.

Example 14.10. For the coarsely triangulated oval plate, the reduced stiffness matrix is (14.49). 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 (14.49), while

$$\mathbf{b} = \tfrac{4}{3}\left(2 + \tfrac{3\sqrt{3}}{4},\ 2,\ 2 + \tfrac{3\sqrt{3}}{4}\right)^T = (\,4.39872, 2.66667, 4.39872\,)^T .$$

The entries of $\mathbf{b}$ are, by (14.52), equal to $4 = f(x_i, y_i)$ times one third the area of the corresponding vertex polygon, which for node 2 is the square consisting of 4 right triangles, each of area $\frac{1}{2}$, whereas for nodes 1 and 3 it consists of 4 right triangles of area $\frac{1}{2}$ plus three equilateral triangles, each of area $\frac{\sqrt{3}}{4}$; see Figure 14.12.

The solution to the final linear system is easily found:

$$\mathbf{c} = (\,1.56724, 1.45028, 1.56724\,)^T .$$

Its entries are the values of the finite element approximation at the three interior nodes. The finite element solution is plotted in the first illustration in Figure 14.15. A more accurate solution, based on a square grid triangulation of size $h = .1$ is plotted in the second figure.

Inhomogeneous Boundary Conditions

So far, we have restricted our attention to problems with homogeneous Dirichlet boundary conditions. According to Theorem 14.5, 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 (14.24). However, now the minimization takes place over the affine subspace consisting of all 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 $\mathbf{x}_k$ is a boundary node, then the corresponding *boundary element* $\varphi_k(x,y)$ satisfies the interpolation condition (14.27), and so has the same piecewise affine form (14.35). The corresponding finite element approximation

$$w(x,y) = \sum_{i=1}^{m} c_i\, \varphi_i(x,y), \tag{14.54}$$

has the same form as before, (14.28), but now the sum is over all nodes, both interior and boundary. As before, the coefficients $c_i = w(x_i, y_i) \approx u(x_i, y_i)$ are the values of the finite element approximation at the nodes. Therefore, in order to satisfy the boundary conditions, we require

$$c_j = h(x_j, y_j) \qquad \text{whenever} \qquad \mathbf{x}_j = (x_j, y_j) \qquad \text{is a boundary node.} \tag{14.55}$$

Remark: If the boundary node $\mathbf{x}_j$ does not lie precisely on the boundary $\partial\Omega$, we need to approximate the value $h(x_j, y_j)$ appropriately, e.g., by using the value of $h(x,y)$ at the nearest 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 the reader, the final outcome can be written as follows. Let 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 the 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 the coarse triangulation of the oval plate, the full finite element matrix is given in (14.48), and the upper 3×3 subblock is the reduced matrix (14.49). The remaining entries of the first three rows form the boundary coefficient matrix

$$\widetilde{K} = \begin{pmatrix} 0 & -.7887 & -.5774 & -.5774 & -.7887 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -.7887 & -.5774 & -.5774 & -.7887 \end{pmatrix}. \tag{14.56}$$

We similarly split the coefficients c_i of the finite element function (14.54) into two groups. We let $\mathbf{c} \in \mathbb{R}^n$ denote the as yet unknown coefficients c_i corresponding to the values of the approximation at the interior nodes $\mathbf{x}_i$, while $\mathbf{h} \in \mathbb{R}^{m-n}$ will be the vector of boundary values (14.55). The solution to the finite element approximation (14.54) is obtained by solving the associated linear system

$$K\,\mathbf{c} + \widetilde{K}\,\mathbf{h} = \mathbf{b}, \qquad \text{or} \qquad K\,\mathbf{c} = \mathbf{f} = \mathbf{b} - \widetilde{K}\,\mathbf{h}. \tag{14.57}$$

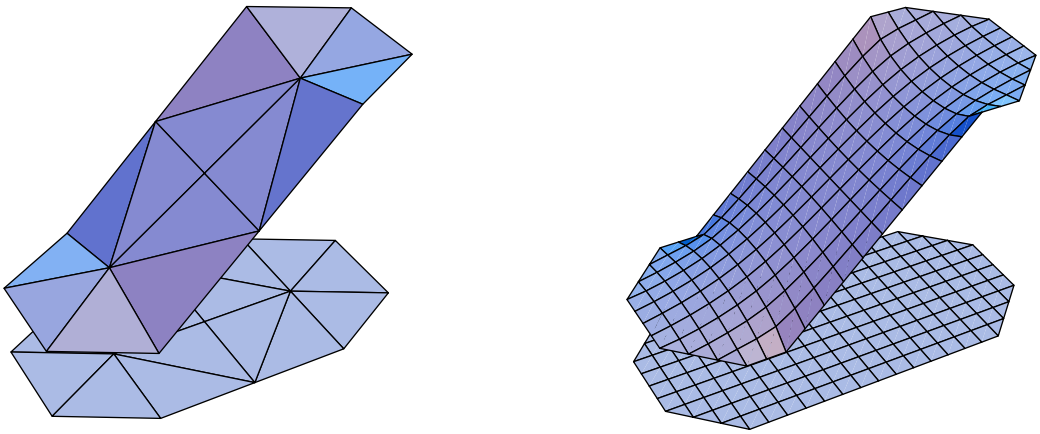

Figure 14.16. Solution to the Dirichlet Problem for the Oval Plate.

Example 14.11. For the oval plate discussed in Example 14.9, 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 14.16. Our task is to compute its equilibrium temperature, assuming no internal heat source. Thus, for the coarse triangulation we have the boundary nodes values

$$\mathbf{h} = (\, h_4, \dots, h_{13} \,)^T = (\, 0, -1, -1, -1, -1, 0, 1, 1, 1, 1, 0 \,)^T .$$

Using the previously computed formulae (14.49, 56) for the interior coefficient matrix K and boundary coefficient matrix $\widetilde{K}$, we approximate the solution to the Laplace equation by solving (14.57). We are assuming that there is no external forcing function, $f(x, y) \equiv 0$, and so the right hand side is $\mathbf{b} = \mathbf{0}$, and so we must solve $K\mathbf{c} = \mathbf{f} = -\widetilde{K}\,\mathbf{h} = (\, 2.18564, 3.6, 7.64974 \,)^T$. The finite element function corresponding to the solution $\mathbf{c} = (\, 1.06795, 1.8, 2.53205 \,)^T$ is plotted in the first illustration in Figure 14.16. 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 a square mesh with spacing $h = .2$ between nodes.

