deciding whether $u(x)$ is p times continuously differentiable we need to look at the function defined by setting

$$u(-x) = -u(x)$$

for $-1 \le x \le 0$ and then extended periodically with period 2 from the interval $[-1, 1]$ to the whole real line. This requires certain properties in u at the endpoints $x = 0$ and $x = 1$. In particular the extended $u(x)$ is C^∞ only if all even derivatives of u vanish at these two points along with u being C^∞ in the interior.

Such difficulties mean that spectral methods based on Fourier series are most suitable in certain special cases (for example if we are solving a problem with periodic boundary conditions in which case we expect the solution to be periodic and have the required smoothness). Methods based on similar ideas can be developed using other classes of functions rather than trigonometric functions and are often used in practice. For example families of orthogonal polynomials such as Chebyshev or Legendre polynomials can be used and fast algorithms developed that achieve spectral accuracy.

4.2.3 Stability

To see that the results quoted above for the local error carry over to the global error as we refine the grid we also need to check that the method is stable. Using the matrix interpretation of the method this is easy to do in the 2-norm. The matrix B in (4.9) is easily seen to be symmetric (recall that $R^{-1} = 2hR = 2hR^T$ and so the 2-norm of B^{-1} is equal to its spectral radius which is clearly $1/\pi^2$ independent of h. Hence the method is stable in the 2-norm.

4.2.4 Collocation property

Though it may not be obvious the approximation we derived above for $U(x)$ in fact satisfies $U''(x_i) = f(x_i)$ at each of the points x_1 through x_m. In other words this spectral method is also a special form of a collocation method as described in Section 4.1.

4.3 The finite element method

The finite element method determines an approximate soution that is a linear combination of some specified basis functions in a very different way from collocation or expansion in eigenfunctions. This method is typically based on some "weak form" of the differential equation which roughly speaking means that we have integrated the equation.

Consider for example the heat conduction problem in one dimension with a variable conductivity $\kappa(x)$ so the steady-state equation is

$$(\kappa u')' = f. \tag{4.10}$$

Again for simplicity assume that the boundary conditions are $u(0) = u(1) = 0$. If we multiply both sides of the equation (4.10) by an arbitrary smooth function $v(x)$ and integrate the resulting product over the domain $[0, 1]$ we obtain

$$\int_0^1 (\kappa(x)u'(x))'v(x)\,dx = \int_0^1 f(x)v(x)\,dx. \tag{4.11}$$

On the left-hand side we can integrate by parts. Since v is arbitrary let's restrict our attention to v that satisfy $v(0) = v(1) = 0$ so that the boundary terms drop out yielding

$$-\int_0^1 \kappa(x)u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx. \tag{4.12}$$

It can be shown that if $u(x)$ satisfies this equation for all v in some suitable class of functions then $u(x)$ is in fact the solution to the original differential equation.

Now suppose we replace $u(x)$ by an approximation $U(x)$ in this expression, where $U(x)$ is a linear combination of specified basis functions,

$$U(x) = \sum_{j=1}^{m} c_j \phi_j(x). \tag{4.13}$$

Let's suppose that our basis functions are chosen to satisfy $\phi_j(0) = \phi_j(1) = 0$, so that $U(x)$ automatically satisfies the boundary conditions regardless of how we choose the c_j. Then we could try to choose the coefficients c_j in $U(x)$ so that the equality (4.12) is satisfied for a large class of functions $v(x)$. Since we only have m free parameters, we can't require that (4.12) be satisfied for all smooth functions $v(x)$, but we can require that it be satisfied for all functions in some m-dimensional function space. Such a space is determined by a set of m basis functions $\psi_i(x)$ (which might or might not be the same as the functions $\phi_j(x)$). If we require that (4.12) be satisfied for the special case where v is chosen to be any one of these functions, then by linearity (4.12) will also be satisfied for any v that is an arbitrary linear combination of these functions, and hence for all v in this m-dimensional linear space.

Hence we are going to require that

$$-\int_0^1 \kappa(x) \left(\sum_{j=1}^{m} c_j \phi_j'(x) \right) \psi_i'(x)\, dx = \int_0^1 f(x)\psi_i(x)\, dx \tag{4.14}$$

for $i = 1,\ 2,\ \ldots,\ m$. We can rearrange this to give

$$\sum_{j=1}^{m} K_{ij} c_j = \int_0^1 f(x)\psi_i(x)\, dx \tag{4.15}$$

where

$$K_{ij} = -\int_0^1 \kappa(x)\phi_j'(x)\psi_i'(x)\, dx. \tag{4.16}$$

The equations (4.15) for $i = 1,\ 2,\ \ldots,\ m$ give an $m \times m$ linear system of equations to solve for the c_j, which we could write as

$$Kc = F$$

with

$$F_i = \int_0^1 f(x)\psi_i(x)\, dx. \tag{4.17}$$

The functions ψ_i are generally called "test functions" while the basis functions ϕ_i for our approximate solution are called "trial functions". Frequently the same basis functions are used for both spaces. The resulting method is known as the *Galerkin method*. If the trial space is different from the test space we have a *Petrov-Galerkin method*.

Example 4.3. As a specific example, consider the Galerkin method for the above problem with basis functions defined as follows on a uniform grid with $x_i = ih$ and $h = 1/(m+1)$. The j'th basis function $\phi_j(x)$ is

$$\phi_j(x) = \begin{cases} (x - x_{j-1})/h & \text{if } x_{j-1} \le x \le x_j \\ (x_{j+1} - x)/h & \text{if } x_j \le x \le x_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{4.18}$$

Each of these functions is continuous and piecewise linear, and $\phi_j(x)$ takes the value 1 at x_j and the value 0 at all other nodes x_i for $i \neq j$. (See Figure 4.1(a).) Note that any linear combination (4.13) of these functions will still be continuous and piecewise linear, and will take the value x_i at the point x_i

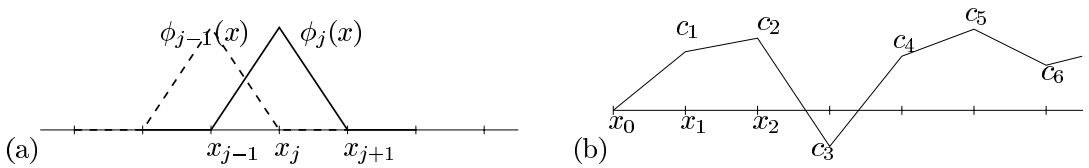

Figure 4.1: (a) Two typical basis functions $\phi_{j-1}(x)$ and $\phi_j(x)$ with continuous piecewise linear elements. (b) $U(x)$Γa typical linear combination such basis functions.

since $U(x_i) = \sum_j c_j\phi_j(x_i) = c_i$ since all other terms in the sum are zero. Hence the function $U(x)$ has the form shown in Figure 4.1(b).

The set of functions $\{\phi_j(x)\}$ form a basis for the space of all continuous piecewise linear functions defined on $[0, 1]$ with $u(0) = u(1) = 0$ and with kinks at the points x_1, x_2, ... , x_mΓwhich are called the *nodes*. Note that the coefficient c_j can be interpreted as the value of the approximate solution at the point x_j.

To use these basis functions in the Galerkin equations (4.14) (with $\psi_j = \phi_j$)Γwe need to compute the derivatives of these basis functions and then the elements of the matrix K and right-hand side F. We have

$$\phi_j'(x) = \begin{cases} 1/h & \text{if } x_{j-1} \le x \le x_j \\ -1/h & \text{if } x_j \le x \le x_{j+1} \\ 0 & \text{otherwise.} \end{cases}$$

For general functions $\kappa(x)$ we might have to compute an approximation to the integral in (4.16)Γbut as a simple example consider the case $\kappa(x) \equiv 1$ (so the equation is just $u''(x) = f(x)$. Then we can compute that

$$K_{ij} = -\int_0^1 \phi_j'(x)\phi_i'(x)\,dx = \begin{cases} 1/h & \text{if } j = i-1 \text{ or } j = i+1, \\ -2/h & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

The matrix K is quite familiar (except for the different power of h):

$$K = \frac{1}{h}\begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{bmatrix}. \tag{4.19}$$

In some cases we may be able to evaluate the integral in (4.17) for F_i explicitly. More generally we might use a discrete approximation. Note that since $\phi_i(x)$ is nonzero only near x_iΓand $\int_0^1 \phi_i(x)\,dx = h$Γ this is roughly

$$F_i \approx hf(x_i). \tag{4.20}$$

In fact the trapezoidal method applied to this integral on the same grid would give exactly this result. Using (4.20) in the system $Kc = F$Γand dividing both sides by hΓgives exactly the same linear system of equations that we obtained in Section 2.4 from the finite difference method (for the case $\alpha = \beta = 0$ we are considering here).

Exercise 4.1 *If we have more general Dirichlet boundary conditions $u(0) = \alpha$ and $u(1) = \beta$, we can introduce two additional basis functions $\phi_0(x)$ and $\phi_{m+1}(x)$ which are also defined by (4.18). Then we know $c_0 = \alpha$ and $c_{m+1} = \beta$ and these terms in the extended sum appearing in (4.12) can be moved to the right hand side. Carry this through to see that we get essentially the system (2.9) in this more general case as well.*

Some comments on this method:

- The matrix K above is tridiagonal because each $\phi_j(x)$ is nonzero on only two elements for $x_{j-1} < x < x_{j+1}$. The function $\phi_i'(x)\phi_j'(x)$ is identically zero unless $j = i-1$, i or $i+1$. More generally, if we choose basis functions that are nonzero only on some region $x_{j-b} < x < x_{j+a}$ then the resulting matrix would be banded with b diagonals of nonzeros below the diagonal and a bands above. In the finite element method one almost always chooses local basis functions of this sort, that are each nonzero over only a few elements.
- Why did we integrate by parts to obtain equation (4.12), rather than working directly with (4.11)? One could go through the same process based on (4.11), but then we would need an approximate $U(x)$ with meaningful second derivatives. This would rule out the use of the simple piecewise linear basis functions used above. (Note that the piecewise linear functions don't have meaningful first derivatives at the nodes, but since only integrals of these functions are used in defining the matrix this is not a problem.)
- This is one advantage of the finite element method over collocation, for example. One can often use functions $U(x)$ for which the original differential equation does not even make sense because U is not sufficiently smooth.
- There are other good reasons for integrating by parts. The resulting equation (4.12) can also be derived from a variational principle and has physical meaning in terms of minimizing the "energy" in the system. (See, *e.g.*, [SF73].)

4.3.1 Two space dimensions

In the last example we saw that the one-dimensional finite element method based on piecewise linear elements is equivalent to the finite difference method derived in Section 2.4. Since it is considerably more complicated to derive via the finite element approach, this may not seem like a useful technique. However, in more than one dimension this method can be extended to irregular grids on complicated regions for which it would not be so easy to derive a finite difference method.

Consider, for example, the Poisson problem with homogeneous Dirichlet boundary conditions on the region shown in Figure 4.2, which also shows a fairly coarse "triangulation" of the region. The points (x_j, y_j) at the corners of the triangles are called *nodes*. The Galerkin form of the Poisson problem is

$$-\int\int_\Omega \nabla q \cdot \nabla v \, dx\, dy = \int\int_\Omega fv \, dx\, dy. \tag{4.21}$$

This should hold for all test functions $v(x, y)$ in some class. Again we can approximate $u(x, y)$ by some linear combination of specified basis functions:

$$U(x, y) = \sum_{j=1}^{N} c_j \phi_j(x, y). \tag{4.22}$$

Taking an approach analogous to the one-dimensional case above, we can define a basis function $\phi_j(x, y)$ associated with each node (x_j, y_j) to be the unique function that is linear on each triangle, and which takes the value 1 at the node (x_j, y_j) and 0 at all other nodes. This function is continuous across the boundaries between triangles and nonzero only for the triangles that have Node j as a corner. For example, Figure 4.2 indicates contour lines for the basis function $\phi_8(x, y)$ as dashed lines.

Using (4.22) in (4.21) gives an $N \times N$ linear system of the form $Kc = F$ where

$$K_{ij} = -\int\int_\Omega \nabla\phi_j \cdot \nabla\phi_i \, dx\, dy. \tag{4.23}$$

These gradients are easy to compute and in fact are constant within each triangle since the basis function is linear there. Since $\nabla\phi_i$ is identically zero in any triangle for which Node i is not a corner, we see

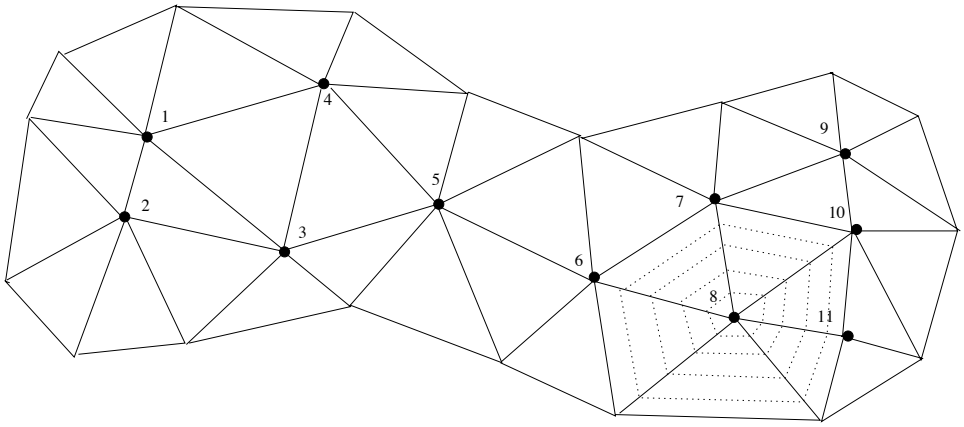

Figure 4.2: Triangulation of a two-dimensional region with 11 nodes. Contourlines for the basis function $\phi_8(x, y)$ are also shown as dashed lines.

that $K_{ij} = 0$ unless Nodes i and j are two corners of a common triangle. For example in Figure 4.2 the eighth row of the matrix K will have have nonzeros only in columns 6, 7, 8, 10 and 11.

Note also that K will be a symmetric matrix since the expression (4.23) is symmetric in i and j. It can also be shown to be positive definite.

For a typical triangulation on a much finer grid we would have a large but very sparse matrix K. The structure of the matrix however will not generally be as simple as what we would obtain with a finite difference method on a rectangular grid. The pattern of nonzeros will depend greatly on how we order the unknowns and equations. Direct methods for solving such systems rely greatly on algorithms for ordering them to minimize the bandwidth. See [DER86].