Chapter 4

Finite Element Method for Ordinary Differential Equations

In this chapter we consider some simple examples of the finite element method for the approximate solution of ordinary differential equations. Although the principal use of finite element methods is for approximating solutions to partial differential equations, it is instructive to look at one-dimensional problems for their simplicity and ease of understanding. In addition, when we approximate PDEs using rectangular elements, then we take tensor products of one-dimensional elements.

In the first three examples we consider a two-point boundary value problem for a second-order linear ordinary differential equation. Each of these examples is constructed so that the approach for handling different boundary data is made evident. The fourth example is a higher order differential equation.

In each example we define an appropriate weak formulation, either prove or indicate how the hypotheses of the Lax-Milgram theorem can be established, discuss the finite element approximation of the weak problem, and present error estimates. In addition, we provide computational results for some examples.

4.1 A two-point BVP with homogeneous Dirichlet boundary data

We begin by considering the following two-point boundary value problem on $[0,1]$ where we seek a function $u(x)$ satisfying

$$\begin{aligned} -\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u &= f(x) \quad \text{for } 0 < x < 1 \\ u(0) &= 0 \\ u(1) &= 0\,, \end{aligned} \tag{4.1}$$

where $p(x)$, $q(x)$, and $f(x)$ are given functions defined on $[0,1]$. In the sequel we assume that $0 < p_{\min} \le p(x) \le p_{\max}$ and $q_{\min} = 0 \le q(x) \le q_{\max}$ where $p_{\min}$, $p_{\max}$, and $q_{\max}$ are positive constants and $f \in L^2(0,1)$. This problem is often referred to as a Sturm-Liouville problem.

It is well-known that whenever $f, q \in C[0,1]$ and $p \in C^1[0,1]$ the boundary value problem (4.1) possesses a unique *classical solution* $u(x) \in C^2(0,1)$ which satisfies (4.1) for every $x \in [0,1]$. We are interested in a *weak* or *generalized solution* of (4.1); *i.e.,* in a function $u(x)$ that satisfies (4.1) in some sense even when f, p, q are not continuous; if f, p, q are sufficiently smooth then we want the weak solution to coincide with the classical solution.

4.1.1 Weak formulation

In choosing the underlying Hilbert space for our weak formulation of (4.1), we know that multiplication of the differential equation by an appropriate test function, integrating over the domain and then integrating by parts to balance the order of the derivatives results in both the test and trial functions having one derivative. Consequently we require our solution to be in $L^2(0,1)$ and to possess at least one weak L^2-derivative. In addition, we constrain our space so that we only consider functions which satisfy the homogeneous Dirichlet boundary conditions. Thus we choose $H_0^1(0,1)$ to be the underlying Hilbert space in which we seek a solution $u(x)$ and for our test space. On $H_0^1(0,1)$ we define the bilinear form $A(\cdot,\cdot)$ by

$$A(v,w) = \int_0^1 p(x)v'(x)w'(x)\,dx + \int_0^1 q(x)v(x)w(x)\,dx = (pv', w') + (qv, w)\ , \quad (4.2)$$

where $(\cdot,\cdot)$ denotes the standard $L^2(\Omega)$-inner product. The weak problem is stated as:

$$\begin{cases} \text{seek } u \in H_0^1(0,1) \text{ satisfying} \\ \qquad A(u,v) = (f,v) \quad \forall\, v \in H_0^1(0,1)\,. \end{cases} \quad (4.3)$$

Note that if u is the classical solution of (4.1) then $u(x)$ also satisfies the weak problem because for $v \in H_0^1(0,1)$

$$\begin{aligned} (f,v) = \int_0^1 fv\,dx &= \int_0^1 \big(-(pu')' + qu\big)v\,dx \\ &= \int_0^1 pu'v'\,dx + \int_0^1 quv\,dx - \big[pu'v\big]\,\big|_0^1 \\ &= \int_0^1 pu'v'\,dx + \int_0^1 quv\,dx = A(u,v)\,. \end{aligned}$$

Conversely, if $u \in H_0^1(0,1)$ satisfies (4.3) and is u is sufficiently smooth, *i.e.,* $u \in C^2(0,1)$, a situation which can be guaranteed if p, q and f are themselves sufficiently smooth, then u coincides with the classical solution of (4.1). The homogeneous Dirichlet boundary conditions are satisfied because $u \in H_0^1(0,1)$ and the differential equation holds because

$$\begin{aligned} A(u,v) - (f,v) &= \int_0^1 pu'v'\,dx + \int_0^1 quv\,dx - \int_0^1 fv\,dx \\ &= \int_0^1 \big[(-pu')' + qu - f\big]v\,dx = 0 \quad \forall\, v \in H_0^1(0,1) \end{aligned}$$

and $v \in H_0^1(0,1)$ is arbitrary. Recall that if we can find a function $u \in H_0^1(0,1)$ which is the unique solution of (4.3), then we call u the *weak solution* of (4.1) in $H_0^1(0,1)$.

To prove the existence and uniqueness of $u \in H_0^1(0,1)$ satisfying (4.3) we use the Lax-Milgram theorem (Theorem **??**) and verify that $A(\cdot,\cdot)$ and $F(v)$ satisfy the hypotheses of this theorem. Clearly, $A(\cdot,\cdot)$ is a bilinear form on $H_0^1(0,1) \times H_0^1(0,1)$. We first show that it is bounded on the space $H_0^1(0,1)$, i.e., $|A(v,w)| \le M \|v\|_1 \|u\|_1$. To do this we use properties of integrals, the given bounds on p, q and the Cauchy-Schwartz inequality to obtain

$$\begin{aligned} |A(v,w)| &\le \Big| \int_0^1 p(x) v' w' \, dx \Big| + \Big| \int_0^1 q(x) v w \, dx \Big| \\ &\le p_{\max} \Big| \int_0^1 v' w' \, dx \Big| + q_{\max} \Big| \int_0^1 v w \, dx \Big| \\ &= p_{\max} \Big| (v', w') \Big| + q_{\max} \Big| (v, w) \Big| \\ &\le p_{\max} \|v'\|_0 \|w'\|_0 + q_{\max} \|v\|_0 \|w\|_0 \, . \end{aligned}$$

To complete the result, we note that by the definition of the L^2-norm and the H^1-norm and seminorm, $\|w'\|_0 = |w|_1$, $\|\cdot\|_0 \le \|\cdot\|_1$, $|\cdot|_1 \le \|\cdot\|_1$. Thus

$$|A(v,w)| \le p_{\max} \|v\|_1 \|w\|_1 + q_{\max} \|v\|_1 \|w\|_1 \le C \|v\|_1 \|w\|_1 \, ,$$

where $C = p_{\max} + q_{\max}$. Therefore, condition (**??**) of the Lax-Milgram theorem is satisfied.

In general, demonstrating coercivity of the bilinear form usually requires more finesse than proving continuity. We must prove that $A(v,v) \ge m \|v\|_1^2$. In our case we have

$$A(v,v) = \int_0^1 p (v')^2 \, dx + \int_0^1 q v^2 \, dx \ge p_{\min} \|v'\|_0^2 + q_{\min} \|v\|_0^2 \, .$$

But we have assumed $q_{\min} = 0$ so

$$A(v,v) \ge p_{\min} \|v'\|_0^2 \, .$$

We must now bound $\|v'\|_0 = |v|_1$ below by a constant times $\|v\|_1$ for all $v \in H_0^1(0,1)$. The fact that $v \in H_0^1(0,1)$ allows us to use the Poincaré inequality (**??**) to bound $|v|_1 \ge \frac{1}{C_p} \|v\|_0$. Using this bound for the entire term $\|v'\|_0^2 = |v'|_1^2$ does not give us the desired result so we use the approach of breaking this term into two parts; we have

$$p_{\min} \|v'\|_0^2 = p_{\min} |v|_1^2 = p_{\min} \Big(\frac{1}{2} |v|_1^2 + \frac{1}{2} |v|_1^2 \Big) \ge \frac{1}{2} p_{\min} \Big(|v|_1^2 + \frac{1}{C_p^2} \|v\|_0^2 \Big) \, .$$

Then

$$A(v,v) \ge \frac{1}{2} p_{\min} \Big[\min \left(1, \frac{1}{C_p^2} \right) \Big] \Big(|v|_1^2 + \|v\|_0^2 \Big) = m \|v\|_1^2 \, ,$$

where we have used the definition of the H^1-norm, $\|\cdot\|_1^2 = \|\cdot\|_0^2 + |\cdot|_1^2$; thus the coercivity condition (**??**) is satisfied. Clearly $F(v) = (f, v)$ is a bounded linear functional on $H_0^1(0,1)$. Thus the Lax-Milgram theorem guarantees the existence of a unique $u \in H_0^1(0,1)$ which satisfies (4.3).

In this problem we constrained our Hilbert space to consist of functions which satisfy the homogenous Dirichlet boundary conditions. We recall that boundary conditions which are satisfied by constraining the admissible or trial space are called *essential*.

4.1.2 Approximation using piecewise linear polynomials

We now turn to approximating u, the solution of the weak problem (4.3), by its Galerkin approximation u^h in a finite dimensional subspace S_0^h of $H_0^1(0,1)$. The approximate solution is required to satisfy (4.3) but only for all $v^h \in S_0^h$; the discrete weak problem is

$$\begin{cases} \text{seek } u^h \in S_0^h \text{ satisfying} \\ \qquad A(u^h, v^h) = \big(f, v^h\big) \quad \forall\, v^h \in S_0^h\,. \end{cases} \tag{4.4}$$

Because $S_0^h \subset H_0^1(0,1)$ the conditions of the Lax Milgram theorem are automatically satisfied on S_0^h and so we are guaranteed that there exists a unique $u^h \in S_0^h$ which satisfies (4.4). Moreover, Galerkin/Cea's Lemma gives us the error estimate

$$\big\|u - u^h\big\|_1 \le C \inf_{\chi^h \in S_0^h} \big\|u - \chi^h\big\|_1\,. \tag{4.5}$$

First we choose S_0^h to be the space of continuous linear piecewise polynomials defined on a partition of $[0,1]$ which satisify the homogeneous Dirichlet boundary conditions. In particular, we consider the following partition of $[0,1]$:

$$0 = x_0 < x_1 < \cdots < x_{N+1} = 1 \quad \text{where} \quad x_i = x_{i-1} + h_i, \quad 1 \le i \le N+1\,, \tag{4.6}$$

and where h_i, $1 \le i \le N+1$ are given numbers such that $0 < h_i < 1$ and $\sum_{i=1}^{N+1} h_i = 1$. We define $h = \max_{1 \le i \le N+1} h_i$; if $h_i = h$ for all i then we call the subdivision *uniform*. A *continuous piecewise linear function* with respect to the given subdivision on $[0,1]$ is a function $\phi(x)$ defined on $[0,1]$ which is linear on each subinterval; *i.e.,* $\phi(x) = \alpha_i x + \beta_i$ on $[x_i, x_{i+1}]$, $0 \le i \le N$. To impose continuity we require that the constants satisfy α_i, β_i where $\alpha_{i-1}x_i + \beta_{i-1} = \alpha_i x_i + \beta_i$, $i = 1, \ldots, N$; We define

$$\begin{aligned} S_0^h = \{\phi(x) \,:\, & \phi \in C[0,1], \\ & \phi(x) \text{ linear on } [x_i, x_{i+1}] \text{ for } 0 \le i \le N, \phi(0) = \phi(1) = 0\}\,. \end{aligned} \tag{4.7}$$

As we discussed in Chapter **??** we want to choose a basis whose functions have as small support as possible so that the resulting coefficient matrix is sparse. For $1 \le i \le N$ we consider again the "hat" functions (see Figure **??**)

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h_i} & \text{for } x_{i-1} \le x \le x_i \\ \dfrac{x_{i+1} - x}{h_{i+1}} & \text{for } x_i \le x \le x_{i+1} \\ 0 & \text{elsewhere.} \end{cases} \tag{4.8}$$

Clearly $\phi_i(x) \in S_0^h$ for $1 \le i \le N$. Moreover, we easily see that

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{4.9}$$

for $1 \le i \le N$ and $0 \le j \le N+1$. Here δ_{ij} denotes the Kronecker delta function. The following proposition justifies our intuition that the functions defined in (4.8) form a basis for S_0^h.

Proposition 4.1. *S_0^h defined by (4.7) is an N-dimensional subspace of $H_0^1(0,1)$. The functions $\{\phi_i(x)\}_{i=1}^N$ defined in (4.8) form a basis for S_0^h.*

Proof. Every function $\phi(x) \in S_0^h$ also belongs to $L^2(0,1)$ and each function is piecewise linear, so analogous to the function $|x|$, each has a weak derivative (which is piecewise constant) in $L^2(0,1)$. Also $\phi(0) = \phi(1) = 0$ for all $\phi \in S_0^h$ so that $S_0^h \subset H_0^1(0,1)$. We now show that $\{\phi_i(x)\}$, $i = 1, \ldots, N$ are linearly independent and span the space S_0^h. To see that we have a linearly independent set, let $\psi(x) = \sum_{i=1}^N c_i \phi_i(x)$; we want to show that the only way $\psi(x) = 0$ for all x is if $c_i = 0$ for $i = 1, \ldots, N$. Using (4.9), we see that $\psi(x_i) = c_i$ for $1 \le i \le N$. Thus if $\psi(x) = 0$ for all x we have that $c_i = 0$ for $i = 1, \ldots, N$; in addition if $c_i = 0$ for all $1 \le i \le N$ then the nodal values of ψ are zero and since it is piecewise linear, it is zero everywhere. Hence we conclude that the functions are linearly independent. To show that the set spans S_0^h we let $\psi(x)$ be an arbitrary element of S_0^h and show that we can write $\psi(x)$ as a linear combination of the $\phi_i(x)$, $i = 1, \ldots, N$; *i.e.,* $\psi(x) = \sum_{i=1}^N c_i \phi_i(x)$. But this can be done by letting $c_i = \psi(x_i)$, *i.e.,* setting c_i to be the *nodal values* of ψ. ∎

Once we have chosen a basis for S_0^h, the problem (4.4) reduces to solving a system of N algebraic equations in N unknowns. Since $u^h \in S_0^h$, we let $u^h(x) = \sum_{j=1}^N \xi_j \phi_j(x)$ and write (4.4) as

$$\sum_{j=1}^{N} \xi_j A(\phi_j, \phi_i) = (f, \phi_i) \quad \text{for } 1 \le i \le N\,.$$

Then $\vec{c} = (\xi_1, \xi_2, \ldots, \xi_N)^T$ satisfies the matrix system

$$\mathcal{A}\vec{c} = \vec{b}\,, \tag{4.10}$$

where $\vec{b} = \left((f,\phi_1), (f,\phi_2), \ldots, (f,\phi_N)\right)^T$ and $\mathcal{A}$ is the $N \times N$ matrix whose elements are given by

$$\mathcal{A}_{ij} = A(\phi_j, \phi_i) = \left(p\phi_j', \phi_i'\right) + (q\phi_j, \phi_i)$$

or

$$\mathcal{A}_{ij} = \mathcal{S}_{ij} + \mathcal{M}_{ij}$$

with $\mathcal{S}_{ij} = \left(p\phi_j', \phi_i'\right)$ and $\mathcal{M}_{ij} = (q\phi_j, \phi_i)$. The matrix $\mathcal{A}$ is symmetric, positive definite (see the exercises) and tridiagonal. If $p(x) = q(x) = 1$ on $[0,1]$ and we use

a uniform mesh, then the matrices $\mathcal{S}$ and $\mathcal{M}$ are explicitly given by

$$\mathcal{S} = \frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & & 0 & -1 & 2 \end{pmatrix} \tag{4.11}$$

and

$$\mathcal{M} = \frac{h}{6}\begin{pmatrix} 4 & 1 & 0 & & \cdots & 0 \\ 1 & 4 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 4 & 1 & 0 & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & 1 & 4 & 1 \\ 0 & \cdots & & 0 & 1 & 4 \end{pmatrix}. \tag{4.12}$$

In the case $p = 1$ the matrix $\mathcal{S}$ is called the *stiffness matrix* of the basis $\{\phi_i\}_{i=1}^N$ while in the case $q = 1$, the matrix $\mathcal{M}$ is called the *Gram matrix* or the *mass matrix* associated with the basis $\{\phi_i\}_{i=1}^N$.

Solution of the linear system

Our coefficient matrix is a symmetric, positive-definite, tridiagonal matrix. If we choose a direct solver, then a Cholesky tridiagonal solver should be used because it takes advantage of these properties of the matrix. Recall that in a Cholesky factorization we write $\mathcal{A} = LL^T$ where L is a lower triangular matrix with positive elements on the diagonal. If $\mathcal{A}$ is the tridiagonal matrix

$$\mathcal{A} = \begin{pmatrix} a_1 & b_2 & & \\ b_2 & a_2 & b_3 & \\ & \ddots & \ddots & \ddots \\ & & b_N & a_N \end{pmatrix} = \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_N & \alpha_N \end{pmatrix} \begin{pmatrix} \alpha_1 & \beta_2 & & \\ & \alpha_2 & \beta_3 & \\ & & \ddots & \ddots \\ & & & \alpha_N \end{pmatrix}$$

then

$$\begin{aligned} \alpha_1 &= \sqrt{a_1} \\ \text{for } i = 2, \ldots, N \qquad \beta_i &= b_i/a_{i-1} \quad \text{and} \quad \alpha_i = \sqrt{a_i - \beta_i^2}\,. \end{aligned} \tag{4.13}$$

Note that we can not determine all the β_i first and then determine the α_i but rather for each i we must determine β_i and then α_i before incrementing i. To solve the system $\mathcal{A}\vec{c} = \vec{f}$ we write $LL^T\vec{c} = \vec{f}$ and solve $L\vec{y} = \vec{f}$ and $L^T\vec{c} = \vec{y}$. Doing this we have the equations

$$\begin{aligned} y_1 &= f_1/\alpha_1 \\ \text{for } i = 2, \ldots, N \qquad y_i &= \frac{f_i - \beta_i y_{i-1}}{\alpha_i} \end{aligned} \tag{4.14}$$

and

$$\begin{aligned} & & c_N &= y_N/\alpha_N \\ \text{for } i = N-1,\ldots,1 & & c_i &= \frac{y_i - \beta_{i+1}y_{i+1}}{\alpha_i}\,. \end{aligned} \tag{4.15}$$

For simplicity of exposition we have defined new variables, α_i, β_i, y_i, and c_i while in practice the entries of $\mathcal{A}$ and $\vec{f}$ are overwritten and no new arrays need be defined.

Error estimates and interpolation results

The bound for the error (4.5) in terms of the error in u and its best approximation in the subspace is not particularly useful in computations; what we would like is to measure the error in terms of powers of h. In order to have a quantitative estimate in terms of powers of h we need to estimate the H^1-error in u and its best approximation in S_0^h but this is difficult to do. However, we note that

$$\inf_{\chi^h \in S_0^h} \left\|u - \chi^h\right\|_1 \leq \left\|u - w^h\right\|_1 \quad \text{for any } w^h \in S_0^h$$

is always true by the definition of the best approximation. So we immediately have

$$\left\|u - u^h\right\|_1 \leq C \left\|u - w^h\right\|_1 \quad \text{for any } w^h \in S_0^h. \tag{4.16}$$

Thus we need to find an element of S_0^h for which an approximation result is available.

Recall from elementary numerical analysis that one way to approximate a function is to use a polynomial interpolant; *i.e.,* to find a polynomial which agrees with the given function or its derivatives at a set of points. One such example is a *Lagrange interpolant* which interpolates given data or function values. Due to the fact that one cannot guarantee that the norm of the difference in the function and the Lagrange interpolating polynomial approaches zero as the degree of the polynomial increases, one often considers piecewise polynomial interpolation. In piecewise Lagrange interpolation we put together Lagrange polynomials of a fixed degree to force them to interpolate the given function values or data. For example, a piecewise linear Lagrange polynomial is a continuous function which is a linear polynomial over each subinterval. Clearly, a piecewise linear Lagrange polynomial over the subdivision of $[0,1]$ given in (4.6) which is zero at $x = 0$ and $x = 1$ is an element of S_0^h.

We state the estimates for the error in a function in $H^1(0,1)$ and its S^h-interpolant where S^h is the space of piecewise linear functions defined over the given partition with no boundary conditions imposed; i.e.,

$$S^h = \{\phi(x) \in C[0,1] \,:\, \phi(x) \text{ linear on } [x_i, x_{i+1}] \text{ for } 0 \leq i \leq N\}\,. \tag{4.17}$$

Then these results also hold for $S_0^h \subset H_0^1(0,1)$. If $v(x)$ is a continuous function on $[0,1]$ then we can find a unique element which agrees with $v(x)$ at each of the points x_i, $i = 0,\ldots,N+1$; we call this element of S^h the S^h-interpolant of v and denote it by $I^h v$. Once we have the standard estimate for the approximation of a function by its piecewise linear Lagrange interpolant measured in the H^1-norm, then, we can

use it in (4.16) to obtain an estimate in terms of powers of h. The following lemma gives standard results for approximating a function by its piecewise linear Lagrange interpolant in the L^2 and H^1 norms; see [Prenter] for details.

Lemma 4.2. *Let $f \in H^1(0,1)$ and $S^h \subset H^1(0,1)$ be defined by (4.17); let $I^h f$ denote the S^h-interpolant of f. Then there exists positive constants C_1, C_2, and C_3, independent of h and f, such that*

$$\left\|f - I^h f\right\|_0 \leq C_1 h \left\|f\right\|_1 . \tag{4.18}$$

In addition, if $f \in H^2(0,1)$ then

$$\left\|f - I^h f\right\|_0 \leq C_2 h^2 \left\|f\right\|_2 \tag{4.19}$$

and

$$\left\|f - I^h f\right\|_1 \leq C_3 h \left\|f\right\|_2 . \tag{4.20}$$

It is important to note that if the solution u to our problem is not smooth enough, *i.e.,* $u \in H^1(0,1)$ and $u \notin H^2(0,1)$, then (4.19) and (4.20) do not hold. In this situation we only have (4.18) and $\left\|u - I^h u\right\|_1 \leq C \left\|u\right\|_1$; the latter implying that there is no convergence in h; *i.e.,* as $h \to 0$, $\left\|u - I^h u\right\|_1$ does not approach zero. We say that the rate of convergence in (4.19) is order h squared, which is quadratic convergence, and denote it $\mathcal{O}(h^2)$; similarly the rate of convergence in (4.20) is $\mathcal{O}(h)$ which is linear convergence. From (4.19) and (4.20) we see a pattern arising that the error in the interpolant measured in the L^2 norm is one order higher than the error measured in the H^1 norm; this is due to the fact that the H^1 norm measures errors in the derivatives as well as the function values.

We can now use Lemma 4.2 to state an estimate for the error in u and u^h measured in the H^1-norm in terms of powers of h. We require $u \in H^2(0,1) \cap H_0^1(0,1)$; note that this can be guaranteed if $f, q, p \in L^2(0,1)$. In this case we get the *optimal* rate; this means that we get the same rate of convergence as $h \to 0$ as the interpolant.

Theorem 4.3. *Let $u \in H^2(0,1) \cap H_0^1(0,1)$ and let u^h be the Galerkin approximation of u in the space S_0^h defined by (4.7); i.e., u^h satisfies (4.4). Then there exists a positive constant C, independent of u, h, or u^h such that*

$$\left\|u - u^h\right\|_1 \leq Ch \left\|u\right\|_2 . \tag{4.21}$$

Proof. The proof is an obvious consequence of (4.20) and (4.16). ∎

It is often the case that we are interested in estimating the error in just the function itself and not its derivatives; in this case we want an estimate for the error in the L^2-norm. From the definition of the L^2- and H^1-norms we immediately have that

$$\left\|u - u^h\right\|_0 \leq \left\|u - u^h\right\|_1 \leq Ch \left\|u\right\|_2 ,$$

the latter inequality holding if $u \in H^2(0,1) \cap H_0^1(0,1)$. However, Lemma 4.2 suggests that we should be able to improve the error to $\mathcal{O}(h^2)$; in addition, computations indicate that $O(h^2)$ is attainable. In order to obtain an optimal L^2-estimate, we must assume sufficient smoothness on u and use a technique known as "Nitsche's trick".

Theorem 4.4. *Let $u \in H^2(0,1) \cap H_0^1(0,1)$ be the solution of (4.3) and let u^h be the Galerkin approximation of u in the space S_0^h defined by (4.7) satisfying (4.4). Then there exists a positive constant C, independent of u, h, or u^h such that*

$$\left\|u - u^h\right\|_0 \leq Ch^2 \left\|u\right\|_2 . \tag{4.22}$$

Proof. Let $e = u - u^h$ and let ψ be the unique function in $H_0^1(0,1)$ (whose existence and uniqueness is guaranteed by the Lax-Milgram theorem) satisfying

$$A(\psi, \phi) = (e, \phi) \quad \forall\, \phi \in H_0^1(0,1) . \tag{4.23}$$

Since $e \in H_0^1(0,1)$ we can set $\phi = e$ in the above expression to obtain

$$\left\|e\right\|_0^2 = (e,e) = A(\psi, e) .$$

Now Galerkin orthogonality for this problem guarantees that $A(u - u^h, v^h) = 0$ for all $v^h \in S_0^h$ and thus $A(e, v^h) = 0$ for all $v^h \in S_0^h$ and we can add this term without impunity. We know that $A(\cdot,\cdot)$ is linear and symmetric so we have

$$\left\|e\right\|_0^2 = A(\psi, e) - A(e, v^h) = A(e, \psi - v^h) \quad \forall\, v^h \in S_0^h .$$

Using the boundedness of the bilinear form gives us

$$\left\|e\right\|_0^2 \leq C \left\|e\right\|_1 \left\|\psi - v^h\right\|_1 \quad \forall\, v^h \in S_0^h .$$

We can use Theorem 4.3 to bound $\left\|e\right\|_1$ by $Ch \left\|u\right\|_2$. If we set v^h to be the S_0^h-interpolant of ψ then if $\psi \in H_0^1(0,1) \cap H^2(0,1)$ the estimate (4.20), along with Theorem 4.3 implies

$$\left\|e\right\|_0^2 \leq Ch^2 \left\|\psi\right\|_2 \left\|u\right\|_2 .$$

From the theory of elliptic partial differential equations one can show that if ψ is the solution to (4.23) and $\psi \in H^2(0,1) \cap H_0^1(0,1)$ then we can bound ψ by the L^2-norm of the data; *i.e.,* $\left\|\psi\right\|_2 \leq C \left\|e\right\|_0$. Substituting this bound for ψ into the above expression gives the desired result from

$$\left\|e\right\|_0^2 \leq Ch^2 \left\|e\right\|_0 \left\|u\right\|_2 .$$

■

It is important to realize that in order to get the optimal estimates in the L^2- and H^1-norms, we must have additional smoothness on our solution. This is a consequence of approximation theory, not an artifact of our finite element analysis.

When we present some numerical simulations, we see that a loss in accuracy occurs if our solution is not smooth enough.

We have now completed our analysis of a finite element solution of (4.1) using continuous, piecewise linear polynomials. Before turning our attention to implementing the method to obtain some numerical results we consider approximating using higher degree polynomials and then remind ourselves how the entries in the matrix and right-hand side of (4.10) are obtained.

4.1.3 Approximation using higher degree polynomials

From the error estimate (4.21) we see that the rate of convergence is linear in the H^1 norm. If we want our calculations to converge at a higher rate, such as quadratically, then we have to choose a higher degree polynomial for our approximating space S_0^h. In this section we give some general results for the error in the interpolating polynomial for a kth degree polynomial and then use these to get optimal error estimates for our problem. We also consider a basis for quadratic polynomials and the structure of the resulting linear system which is no longer tridiagonal as it was when we used linear polynomials. The case of continuous, cubic polynomials is left to the exercises.

We now define S^h to be the space of continuous, piecewise polynomials of degree k or less over the partition of $[0,1]$ defined in (4.6), i.e.,

$$S^h = \{\phi(x) \ : \ \phi \in C[0,1], \phi(x) \text{ polynomial of degree} \leq k \text{ on } [x_i, x_{i+1}] \text{ for } 0 \leq i \leq N\}\,. \tag{4.24}$$

S_0^h is defined in the same way except we require $\phi(x)$ to be zero at the endpoints;

$$S_0^h = \{\phi(x) \ : \ \phi \in C[0,1], \phi(x) \text{ polynomial of degree} \leq k \text{ on } [x_i, x_{i+1}] \text{ for } 0 \leq i \leq N, \phi(0) = \phi(1) = 0\}\,. \tag{4.25}$$

A theorem for the S^h-interpolant of functions in H^1 is provided in the following lemma.

Lemma 4.5. *Let $f \in H^{k+1}(0,1)$ and $S^h \subset H^1(0,1)$ where S^h is defined by (4.24); let $I^h f$ denote the S^h-interpolant of f. Then there exists positive constants C_1, C_2, independent of h and f, such that*

$$\left\|f - I^h f\right\|_0 \leq C_1 h^{k+1} \left\|f\right\|_{k+1} \tag{4.26}$$

and

$$\left\|f - I^h f\right\|_1 \leq C_2 h^{k} \left\|f\right\|_{k+1}\,. \tag{4.27}$$

Note that (4.26) reduces to (4.19) and (4.27) reduces to (4.20) when $k = 1$. These are the best rates of convergence possible with a kth degree polynomial. If f is not in $H^{k+1}(0,1)$ then there is a loss in the rates of convergence. For example, if $f \in H^k(0,1)$ and not in $H^{k+1}(0,1)$, then a power of h is lost in each estimate. If

our finite element solution is in $H^{k+1}(0,1)$ then optimal rate of convergence in the H^1 norm are given in the following theorem.

Theorem 4.6. *Let $u \in H^{k+1}(0,1) \cap H_0^1(0,1)$ be the solution of (4.3) and let u^h be the Galerkin approximation of u in the space S_0^h defined by (4.25) satisfying (4.4). Then there exists a positive constant C, independent of u, h, or u^h such that*

$$\left\|u - u^h\right\|_1 \leq Ch^k \left\|u\right\|_{k+1} . \tag{4.28}$$

We note that this estimate says that if the solution is sufficiently smooth, then increasing the degree of the polynomial by one increases the rate of convergence by one.

As before, we are often interested in the L^2 norm of the error. We can mimic the proof of Theorem 4.4 to get the following result when S_0^h is defined by (4.25). See the exercises for details.

Theorem 4.7. *Let $u \in H^{k+1}(0,1) \cap H_0^1(0,1)$ be the solution of (4.3) and let u^h be the Galerkin approximation of u in the space S_0^h defined by (4.25) satisfying (4.4). Then there exists a positive constant C, independent of u, h, or u^h such that*

$$\left\|u - u^h\right\|_0 \leq Ch^{k+1} \left\|u\right\|_{k+1} . \tag{4.29}$$

We note that the optimal rate of convergence in the L^2 norm is one power of h higher than in the H^1 norm which measures the error in the derivatives of the solution as well as the solution itself.

We now turn to the concrete problem of finding a basis for S^h or S_0^h when we choose quadratic polynomials, i.e., $k = 2$. In this case we know that the rates of convergence are $\mathcal{O}(h^2)$ in the H^1 norm and $\mathcal{O}(h^3)$ in the L^2 norm, if the solution is sufficiently smooth. We use the same partition of $[0,1]$ as before, i.e., that given in (4.6). The problem now is that over each element $[x_{i-1}, x_i]$ the basis function must be a quadratic; however, it takes three points to uniquely determine a quadratic. To this end, we add a node in each subinterval; the easiest thing to do is add a node at the midpoint of each subinterval, $x_{i-\frac{1}{2}} = (x_{i-1} + x_i)/2$. We still have $N+1$ elements, but now have the $N+2$ points from the endpoints of the intervals plus the $N+1$ midpoints giving a total of $2N+3$ points. Analogous to the continuous, piecewise linear case, we expect that a basis for S^h for $k = 2$ consists of $2N+3$ elements and for S_0^h we don't need the endpoints so we have $2N+1$ elements in basis.

For simplicity of exposition, we renumber our $2N+3$ nodes as x_i, $i = 0, \ldots, 2N+2$. However, we must remember that the elements are $[x_{2j-2}, x_{2j}]$ for $j = 1, \ldots, N+1$. To determine a nodal basis for S^h we require each ϕ_i in the basis to have the property that it is one at node x_i and zero at all other nodes. In the basis for piecewise linear polynomials we were able to make the support of the basis functions to be two adjacent elements; the same is true in this case. However, now we have two different formulas for the basis functions determined by whether the function is centered at an endpoint of an interval or the midpoint.

To easily get an idea what these quadratic functions look like, we first write the polynomials on $[-1,1]$ with nodes $x=-1,0,1$; we can then translate them to the desired interval. From these we can determine the shape of our basis functions. For the quadratic function which is one at the midpoint, i.e., $x=0$, and zero at $x=\pm 1$ we have $\phi(x)=1-x^2$. For the quadratic function which is one at $x=-1$ and zero at $x=0,1$ we have $\phi(x)=\frac{1}{2}(x^2-x)$. Similarly for a quadratic function which is one at $x=1$ and zero at $x=-1,0$ we have $\phi(x)=\frac{1}{2}(x^2+x)$. These functions are illustrated in Figure 4.1 and have the same shape as the ones on $[x_{2j-2},x_{2j}]$. We can splice together the two functions centered at the endpoints of the interval to get a complete picture of the basis function centered at an endpoint which has support over two intervals; this is demonstrated in the right plot in Figure 4.1. Note that analogous to the case of continuous piecewise linear polynomials the quadratic basis functions will be in C^0 but not C^1.

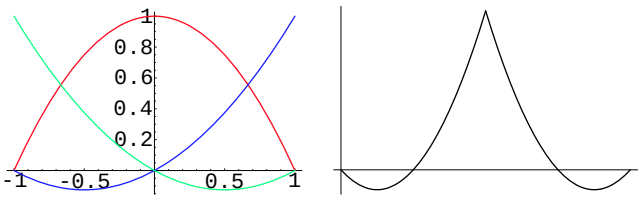

Figure 4.1. *Plot on left shows nodal quadratic functions on* $[-1,1]$ *and plot on right shows shape of quadratic basis function centered at endpoint of an interval having support over two intervals.*

To find the analogous polynomials on $[x_{2j-2},x_{2j}]$ we need to translate our functions on $[-1,1]$ to the desired interval or equivalently solve linear systems. For example, a straightforward way to find the quadratic which is one at x_{2j-1} and zero at the endpoints is to solve

$$\begin{aligned} 0 &= a+b\big(x_{2j-2}\big)+c\big(x_{2j-2}\big)^2 \\ 1 &= a+b\big(x_{2j-1}\big)+c\big(x_{2j-2}\big)^2 \\ 0 &= a+b\big(x_{2j}\big)+c\big(x_{2j}\big)^2 . \end{aligned}$$

In later chapters we discuss more efficient approaches to finding basis functions. The support of the quadratic basis functions for S_0^h on a uniform partition of $[0,2]$ with $h=0.5$ are illustrated in Figure 4.2.

We have seen that once a basis for the finite dimensional space is chosen, the discrete problem can be converted to solving a linear system of equations. The (i,j) entry of the coefficient matrix $\mathcal{A}$ is given by the same expression as in the case of piecewise linear functions except we are using a different basis; specifically, we have

$$\mathcal{A}_{ij}=(p\phi_j',\phi_i')+(q\phi_j,\phi_i)$$

where ϕ_i is now a quadratic polynomial. We recall that when the standard "hat" functions were used as a basis for S_0^h the resulting matrix was $N\times N$, symmetric,

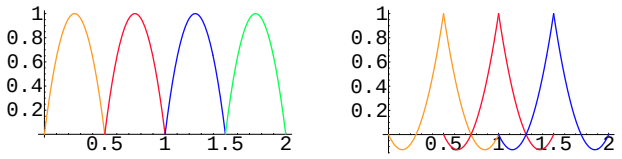

Figure 4.2. *Support of quadratic basis functions on the uniform partition of* $[0, 2]$ *with* $h = .5$ *assuming homogeneous Dirichlet boundary conditions.*

positive definite and tridiagonal. In the case of our quadratic basis functions in S_0^h, the matrix is still symmetric and positive definite but we note that the size of our matrix has increased to $2N + 1$. Also, it is no longer tridiagonal. To determine the bandwidth of the matrix, we need to ascertain where the zero entries begin in each row. We return to Figure 4.2 and note that for a basis function ϕ_i centered at a midpoint node x_i, the integral $\int_0^1 \phi_i \phi_j \, dx$ is zero when $j > i + 1$ or $j < i - 1$, i.e., outside of the interval; the same is true for the term $\int_0^1 \phi_i' \phi_j' \, dx$. However, for a basis function ϕ_i centered at the right endpoint node x_i, the integral $\int_0^1 \phi_i \phi_j \, dx$ is potentially nonzero in that interval and the next which includes a total of five basis functions, counting itself. Thus the integral is zero when $j > i + 2$ or $j < i - 2$ and the maximum bandwidth of the matrix is five. This system can be efficiently solved by a direct method such as a banded Cholesky algorithm or an iterative method such as conjugate gradient or one of its variants.

If we desire to have a method which converges cubically in the H^1 norm, then we can choose continuous, piecewise cubic polynomials for S^h. Because we need four points to uniquely determine a cubic, we add two points to each interval in our original partition given in (4.6). For S_0^h we now have $N + 2(N + 1) = 3N + 2$ points and we expect that this is the dimension of the space and thus the dimension of the resulting matrix. The shape of the basis functions and the structure of the resulting matrix is explored in the exercises.

4.1.4 Numerical quadrature

If we are implementing our example given in (4.1) in the case $p = q = 1$ with continuous, piecewise linear polynomials for S_0^h and where we are using a uniform grid, then (4.11) and (4.12) explicitly give the coefficient matrices. However, entries in the right-hand side of (4.10) must be computed and also entries for the coefficient matrix for general p, q. For some choices of f we could evaluate the integrals exactly. However, if we want to write a general finite element program then we should be able to do problems where the integrals can not be evaluated exactly. In this case, we must use *quadrature rules* to approximate the integrals. Recall that in our error analysis, we have assumed that the integrals are computed exactly; the effects of numerical integration are discussed in a later chapter. For now, we present some widely used quadrature formulas in one-dimension and give general rules for choosing a formula.

In numerical integration we approximate the integral by the sum of the integrand evaluated at a prescribed set of points multiplied by weights; *i.e.,*

$$\int_a^b f(x)\,dx \approx \sum_k f(q_k)w_k\,, \tag{4.30}$$

where q_k represent the quadrature points and w_k the quadrature weights. Of particular interest in one dimension are the Gauss quadrature rules; in these rules the quadrature points and weights are chosen so that the rule integrates exactly as high a degree polynomial as possible. Specifically, if we use n Gaussian quadrature points then the rule integrates polynomials of degree $2n-1$ exactly. The Gaussian quadrature rule for one point is the well known midpoint rule. The following table gives the Gaussian quadrature points and weights on the interval $[-1,1]$. If

Table 4.1. *Gauss quadrature formulas on* $[-1,1]$

n	nodes	weights
1	0.0000000000	2.0000000000
2	±0.5773502692	1.0000000000
3	±0.7745966692	0.5555555556
	0.0000000000	0.8888888889
4	±0.8611363116	0.3478548451
	±0.3399810436	0.6521451549
5	±0.9061798459	0.2369268850
	±0.5384693101	0.4786286701
	0.0000000000	0.5688888889

the domain of integration is different from $(-1,1)$, then a change of variables is needed. For example, to compute the integral $\int_a^b f(\hat{x})\,d\hat{x}$ we use the linear mapping $\hat{x} = a + \frac{b-a}{2}(x+1)$ to map to the integral over $(-1,1)$. In this case we have

$$\int_a^b f(\hat{x})\,d\hat{x} = \frac{b-a}{2}\int_{-1}^1 f(a+\frac{b-a}{2}(x+1)\,dx\,.$$

Then we apply the quadrature rule to the integral over $(-1,1)$. Note that we have just modified the quadrature weight by multiplying by $\frac{b-a}{2}$ and mapping the quadrature point to the interval (a,b).

When choosing a quadrature rule, we want to use as low a degree rule as possible for efficiency but as high a degree rule as necessary for accuracy. It is not necessary to evaluate the integrals exactly, even if this is possible; however, we must assure that the error in the numerical quadrature does not contaminate the power of h accuracy in our estimate. When using piecewise linear polynomials for the finite element space in one-dimension for the problem (4.1), it is adequate to use a one-point Gauss quadrature rules; for piecewise quadratic polynomials a two-point rule is adequate.

4.1.5 Computational examples

In this section we implement two specific examples of the boundary value problem given in (4.1) where we know the exact solution so that errors and rates of convergence can be calculated. These problems differ in the choice of p, q and f. The choice of f is especially important because a lack of smoothness in f results in the solution not being smooth enough to guarantee the optimal rates of convergence. In all computations we use continuous, piecewise polynomials on a uniform grid, an appropriate Gauss quadrature rule to evaluate the integrals in the coefficient matrix and the right-hand side, and a direct solver for the linear system. For the error computation we use a higher order quadrature rule to evaluate the integrals. The reason for the higher order rule in the error computation is to make absolutely sure that no error from the numerical integration contaminates the calculation of the error. The computations are performed using $h = 1/4, 1/8, 1/16$, and $1/32$ with linear, quadratic and cubic elements; the H^1- and L^2-errors are computed for each grid.

For each example we are interested in calculating the numerical rate of convergence and comparing it with the theoretical results presented in Theorems 4.3, 4.4,**??** and **??**. The errors for each grid can be used to compute an approximate rate of convergence. For example, we have $\|u - u^h\| \approx Ch^r$ where we expect r to approach some value as the grid size decreases. If we have the error, E_i, on two separate meshes then we have that $E_1 \approx Ch_1^r$ and $E_2 \approx Ch_2^r$ where E_1 and E_2 represent $\|u - u^h\|$ on the grid with mesh spacing h_1 and h_2, respectively. If we solve for C and set the two relationships equal, we have $E_1/h_1^r \approx E_2/h_2^r$; solving for r we obtain

$$r \approx \frac{\ln E_1/E_2}{\ln h_1/h_2}\,. \tag{4.31}$$

We note that if the grid spacing is halved, i.e., $h_2 = h_1/2$ then the error should be approximately decreased by $\left(\frac{1}{2}\right)^r$ since $E_2 \approx \left(\frac{h_2}{h_1}\right)^r E_1$. This implies that if $r = 1$ the error is approximately halved when the grid spacing is halved; if the rate is two, then the error is reduced by a factor of one-fourth when the grid spacing is halved, etc.

Example 4.8 We first consider the problem

$$\begin{aligned} -u'' + \pi^2 u &= 2x\pi^2 \sin \pi x - 2\pi \cos \pi x \quad \text{for } 0 < x < 1 \\ u(0) = u(1) &= 0\,, \end{aligned} \tag{4.32}$$

whose exact solution is given by $u = x \sin \pi x$. Since our solution $u(x) = x \sin \pi x$ is actually in $C_0^\infty(0,1)$ we expect the optimal rates of convergence; in particular if we use continuous, piecewise linear polynomials then the rate r, calculated from (4.31), should approach two as $h \to 0$ for the L^2-norm and approach one for the H^1-norm. These values for r are calculated in Table 4.2 along with the errors and rates using continuous, piecewise quadratic and cubic polynomials; in the table we computed the rate using the errors at $h = 1/4$ and $1/8$, at $h = 1/8$ and $1/16$, and at $h = 1/16$ and $h = 1/32$. Note that, in fact, $r \to 1$ in the H^1 error and $r \to 2$ in the L^2-error as Theorems 4.3 and 4.4 predict when piecewise linear polynomials

are used; the optimal rates for quadratic and cubic polynomials are also obtained. In these calculations we used a one-point Gauss rule for linear polynomials, a two-point Gauss rule for quadratic polynomials, and a three-point Gauss rule for cubic polynomials. In Table 4.3 we illustrate what happens if we use continuous quadratic polynomials using a one-point, a two-point and a three-point Gauss quadrature rule. Note that the rates of convergence using a two-point and a three-point rule are essentially the same, but when we use the one-point rule the results are meaningless. ∎

Table 4.2. *Numerical results for Example 4.8 using continuous, piecewise linear polynomials.*

p^k	h	$\|u-u^h\|_1$	rate	$\|u-u^h\|_0$	rate
linear	1/4	0.47700		0.28823×10^{-1}	
linear	1/8	0.23783	1.0041	0.69831×10^{-2}	2.0459
linear	1/16	0.11885	1.0007	0.17313×10^{-2}	2.0120
linear	1/32	0.059416	1.0002	0.43199×10^{-3}	2.0028
quadratic	1/4	0.49755×10^{-1}		0.15707×10^{-2}	
quadratic	1/8	0.12649×10^{-1}	1.9758	0.20227×10^{-3}	2.9570
quadratic	1/16	0.31747×10^{-2}	1.9940	0.25553×10^{-4}	2.9847
quadratic	1/32	0.79445×10^{-3}	1.9986	0.32031×10^{-5}	2.9960
cubic	1/4	0.51665×10^{-2}		0.10722×10^{-3}	
cubic	1/8	0.64425×10^{-3}	3.003	0.67724×10^{-5}	3.985
cubic	1/16	0.80496×10^{-4}	3.001	0.42465×10^{-6}	3.9953
cubic	1/32	0.10061×10^{-4}	3.000	0.26564×10^{-7}	3.9987

Example 4.9 The next problem we want to consider is

$$\begin{aligned} -u'' &= -\alpha(\alpha-1)x^{\alpha-2} \quad \text{for } 0<x<1 \\ u(0)=u(1) &= 0\,, \end{aligned} \tag{4.33}$$

where $\alpha > 0$; the exact solution u is given by $u(x) = x^\alpha - x$. The results for various values of α are presented in Table 4.4 using continuous, piecewise linear polynomials and a one-point Gauss quadrature rule. Recall that the optimal rates in this case are $\mathcal{O}(h)$ in the H^1 norm and $\mathcal{O}(h^2)$ in the L^2 norm. Note that for $\alpha = 7/3$ we get the optimal rates of convergence. However, for $\alpha = 4/3$ we have less than optimal rates and for $\alpha = 1/3$ the H^1-error is almost constant and the rate in the L^2-norm is less than one. Of course, the reason for this is that when $\alpha = 3/2$ the exact solution $u = x^{4/3} - x \notin H^2(0,1)$ and when $\alpha = 1/3$ the exact solution $u = x^{1/3} - x \notin H^1(0,1)$. Thus the interpolation results (4.19) and (4.20) do *not* hold and hence Theorems 4.3 and 4.4 do not apply. ∎

Table 4.3. *Numerical results for Example 4.8 using continuous, piecewise quadratic polynomials with three different quadrature rules.*

Gauss Quadrature Rule	h	$\|u-u^h\|_1$	rate	$\|u-u^h\|_0$	rate
one-point	1/4	8.885		0.3904	
one-point	1/8	18.073		0.3665	
one-point	1/16	36.391		0.3603	
one-point	1/32	72.775		0.3587	
two-point	1/4	0.49755×10^{-3}		0.15707×10^{-4}	
two-point	1/8	0.12649×10^{-3}	1.9758	0.20227×10^{-5}	2.9570
two-point	1/16	0.31747×10^{-4}	1.9940	0.25553×10^{-6}	2.9847
two-point	1/32	0.79445×10^{-5}	1.9986	0.32031×10^{-7}	2.9960
three-point	1/4	0.49132×10^{-3}		0.18665×10^{-4}	
three-point	1/8	0.12109×10^{-3}	1.9620	0.24228×10^{-5}	2.9456
three-point	1/16	0.31724×10^{-4}	1.9911	0.30564×10^{-6}	2.9868
three-point	1/32	0.79430×10^{-5}	1.9978	0.38292×10^{-7}	2.9967

Table 4.4. *Numerical results for Example 4.9.*

α	h	$\|u-u^h\|_1$	rate	$\|u-u^h\|_0$	rate
7/3	1/4	0.1747		0.17130×10^{-1}	
7/3	1/8	0.08707	1.0046	0.33455×10^{-2}	1.9726
7/3	1/16	0.04350	1.0012	0.84947×10^{-3}	1.9776
7/3	1/32	0.02174	1.0007	0.21495×10^{-3}	1.9826
4/3	1/4	0.47700		0.28823×10^{-1}	
4/3	1/8	0.23783	0.7690	0.69831×10^{-2}	1.8705
4/3	1/16	0.11885	0.7845	0.17313×10^{-2}	1.8834
4/3	1/32	0.059416	0.7965	0.43199×10^{-3}	1.8005
1/3	1/4	0.43332		0.14594	
1/3	1/8	0.43938		0.10599	0.4615
1/3	1/16	0.46661		0.07922	0.4200
1/3	1/32	0.50890		0.06064	0.3857

4.2 A two-point BVP with Neumann boundary data

In this section we consider the same differential equation as in the first section but now we impose Neumann boundary data instead of homogeneous Dirichlet data. In particular we seek a function $u(x)$ satisfying

$$\begin{aligned} -\frac{d}{dx}\left(p(x)\frac{du}{dx}\right)+q(x)u &= f(x) \quad \text{for } 0<x<1 \\ u'(0) &= 0 \\ u'(1) &= \alpha\,. \end{aligned} \tag{4.34}$$

As before, p and q are bounded functions on $[0,1]$ satisfying $0 < p_{\min} \le p(x) \le p_{\max}$ but now we impose $0 < q_{\min} \le q(x) \le q_{\max}$ for all $x \in [0,1]$. Again if $f, q \in C[0,1]$ and $p \in C^1[0,1]$ the boundary value problem (4.34) possesses a unique *classical solution* $u(x) \in C^2(0,1)$ which satisfies (4.34) for every $x \in [0,1]$. Note that here we require that $q_{\min} > 0$ to guarantee a unique solution; this is because if $q = 0$ and u satisfies (4.34) then so does $u + C$ for any constant C.

In this case our underlying finite element space is $H^1(0,1)$ because we have no boundary conditions to impose on the space. The weak formulation is

$$\begin{cases} \text{seek } u \in H^1(0,1) \text{ satisfying} \\ \qquad A(u,v) = (f,v) + \alpha p(1) v(1) \quad \forall\, v \in H^1(0,1)\,, \end{cases} \tag{4.35}$$

where

$$A(v,w) = \int_0^1 p(x) v'(x) w'(x)\, dx + \int_0^1 q(x) v(x) w(x)\, dx \quad \forall\, v, w \in H^1(0,1)\,.$$

Clearly, if $u(x)$ satisfies the classical problem (4.34), then $u(x)$ satisfies (4.35) because

$$\begin{aligned} \int_0^1 f(x) v\, dx &= \int_0^1 \big(-(p(x)u'(x))' + q(x)u(x)\big) v(x)\, dx \\ &= -pu'v\big|_0^1 + \int_0^1 p(x)u'(x)v'(x)\, dx + \int_0^1 q(x)u(x)v(x)\, dx \\ &= -p(1)u'(1)v(1) + p(0)u'(0)v(0) + A(u,v) \\ &= A(u,v) - \alpha p(1) v(1)\,, \end{aligned}$$

where we have imposed the homogenous Neumann boundary condition $u'(0) = 0$ and the inhomogeneous condition $u'(1) = \alpha$. Note that these boundary conditions are *not* imposed on the space, but rather on the weak formulation; these are called *natural* boundary conditions.

In a manner similar to the example in Section 4.1, we can show that the hypotheses of the Lax-Milgram theorem are satisfied. Recall that in proving coercivity for the previous example, we used the Poincaré inequality to relate the L^2 norm with the H^1 seminorm. We can not longer do this because our function is not zero on any portion of the boundary. However, coercivity can be proved in a straightforward manner; the details are left to the exercises. Thus we are guaranteed the existence and uniqueness of a solution to (4.35).

If we want to seek an approximation to $u(x)$ in the space of continuous, piecewise linear functions defined over the subdivision (4.6) then we cannot use the space S_0^h defined in (4.7) since this space was designed to approximate functions in $H_0^1(0,1)$. Instead we consider S^h where

$$S^h = \{\phi(x) \in C[0,1], \phi(x) \text{ linear on } (x_i, x_{i+1}) \text{ for } 0 \le i \le N\}\,. \tag{4.36}$$

Similar to the homogeneous Dirichlet case, it can be shown that S^h is an $N+2$ dimensional subspace of $H^1(0,1)$; a basis for S^h is given by the standard "hat"

functions that we used for S_0^h along with one defined at each endpoint. Specifically, we have the functions ψ_i, $i = 1, \ldots, N+2$ defined by

$$\psi_i(x) = \begin{cases} \phi_0(x) & \text{for } j = 1 \\ \phi_{i-1}(x) & \text{for } 2 \le i \le N+1 \\ \phi_{N+1}(x) & \text{for } j = N+2 \end{cases} \tag{4.37}$$

where $\phi_i(x)$, $i = 1, \ldots N$ are given by (4.8) and

$$\phi_0(x) = \begin{cases} \dfrac{x_1 - x}{h_1} & \text{for } 0 \le x \le x_1 \\ 0 & \text{elsewhere} \end{cases} \tag{4.38}$$

and

$$\phi_{N+1}(x) = \begin{cases} \dfrac{x - x_N}{h_{N+1}} & \text{for } x_N \le x \le 1 \\ 0 & \text{elsewhere}\,. \end{cases} \tag{4.39}$$

Galerkin's theorem guarantees that there is a unique $u^h \in S^h \subset H^1(0,1)$ satisfying

$$A(u^h, v^h) = (f, v^h) + \alpha p(1) v^h(1) \quad \forall\, v^h \in S^h\,. \tag{4.40}$$

The problem of finding a $u^h \in S^h$ which satisfies (4.40) reduces to solving a linear system of equations; in this case the coefficient matrix has dimension $N+2$. In addition, we can use the interpolation results given in Lemma 4.2 to obtain the following optimal error estimates. See the exercises for a proof.

Theorem 4.10. *Let $u \in H^2(0,1)$ be the solution of (4.34) and let u^h be the Galerkin approximation in S^h defined by (4.36) given by (4.40). Then for some constant C, independent of h, u, and u^h we have*

$$\left\| u - u^h \right\|_k \le C h^{2-k} \left\| u \right\|_2$$

for $k = 0, 1$.

One purpose of the following computations is to demonstrate the difference in satisfying a boundary condition by imposing it on the space (an essential boundary condition) and imposing it weakly through the weak formulation (a natural boundary condition).

Example 4.11 We consider the problem

$$\begin{aligned} -u'' + \pi^2 u &= 2x\pi^2 \sin \pi x - 2\pi \cos \pi x \quad \text{for } 0 < x < 1 \\ u'(0) &= 0 \\ u'(1) &= -\pi\,, \end{aligned} \tag{4.41}$$

whose exact solution is given by $u = x \sin \pi x$. Note that this is the same differential equation as in Example 4.8 but now we are imposing Neumann boundary conditions. Since our solution $u(x) = x \sin \pi x$ is actually in $C^\infty(0,1)$ we expect the optimal rates

of convergence which we can see are obtained from Table 4.5. The approximate solutions using uniform grids of $h = \frac{1}{4}$, $\frac{1}{8}$ and $\frac{1}{16}$ along with the exact solution are plotted in Figure 4.3. Note that although our exact solution is zero at the endpoints, our approximate solution is not because we imposed Neumann boundary conditions. However, the approximate solution does not satisfy the exact derivative boundary condition because we have satisfied it weakly. In the last plot in Figure 4.3 we have blown up the approximate solutions at the right end point which should have a slope of $-\pi$. The approximate derivative at the right boundary is -1.994, -2.645, -2.917 and -3.036 for $h = 1/4$, $1/8$, $1/16$, and $1/32$ respectively. These correspond to errors of 1.147, 0.4968, 0.2244 and 0.1055. As $h \to 0$ the derivative of the approximate solution at $x = 1$ approaches the exact value of $-\pi$ linearly; this is expected because the rate of convergence in the H^1 norm is one. Note that this is in contrast to Example 4.8 where our approximate solution exactly satisfied the homogeneous Dirichlet boundary condition because we imposed it on our space. ∎

Table 4.5. *Numerical results for Example 4.11 using continuous, piecewise linear polynomials.*

	h	$\|u - u^h\|_1$	rate	$\|u - u^h\|_0$	rate
1/4	0.48183		0.22942×10^{-1}		
1/8	0.23838	1.0153	0.56235×10^{-2}	2.0281	
1/16	0.11892	1.0033	0.13988×10^{-2}	2.0073	
1/32	0.059425	1.0009	0.34924×10^{-3}	2.0019	

4.3 A two-point BVP with inhomogeneous boundary data

In the previous two sections we considered two-point boundary values problems with homogeneous Dirichlet boundary data and homogeneous and inhomogeneous Neumann data. Consequently, the only type of boundary conditions that are left to see how to handle are inhomogeneous Dirichlet data and mixed, or Robin, boundary conditions. In this section we demonstrate how an inhomogeneous Dirichlet boundary condition can be handled; the mixed boundary condition is handled similarly to the inhomogeneous Neumann boundary condition. In particular we seek a function $u(x)$ satisfying

$$\begin{aligned} -\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u &= f(x) \quad \text{for } 0 < x < 1 \\ u(0) = \alpha \qquad u'(1) + \sigma u(1) &= \beta\,, \end{aligned} \tag{4.42}$$

where α, β, and σ are constants. Note that if we choose $\sigma = 0$ then we just have an inhomogeneous Neumann boundary condition at the right endpoint as we did in (4.34).

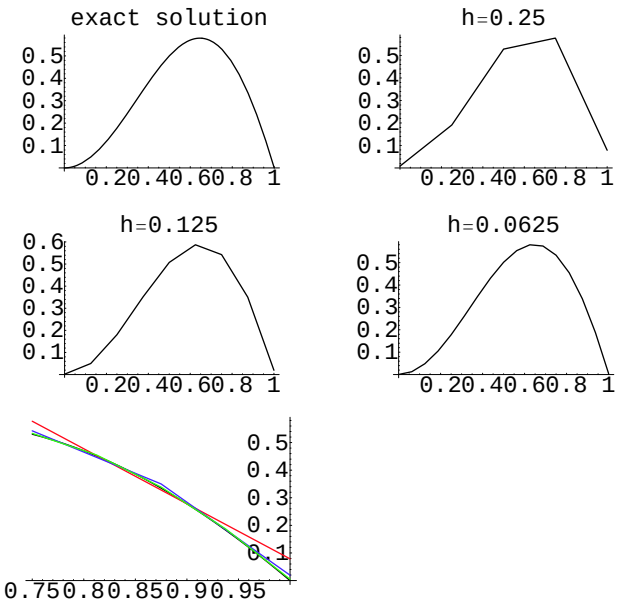

Figure 4.3. *Plots of the exact solution and three piecewise linear approximations. The last plot gives a blow-up of the right endpoint demonstrating that the natural boundary condition is only satisfied weakly.*

We know that the underlying Hilbert space for the weak formulation should be $H^1(0,1)$ or some subspace.. For the Dirichlet boundary condition in Section 4.1 we imposed the boundary condition on the space; *i.e.,* we sought our solution in the subspace of $H^1(0,1)$ consisting of all functions that were zero on the boundary. However, we can not constrain our space to be all functions $\phi \in H^1(0,1)$ which satisfy $\phi(0) = \alpha$. The reason is that this is *not* a subspace of $H^1(0,1)$ since, for example, if $v(0) = \alpha$ and $w(0) = \alpha$ then $(v+w)(0) = 2\alpha$.

Inhomogeneous Dirichlet boundary conditions can be handled in several ways. One of the easiest ways to handle them *theoretically* is to transform the problem into one which has homogeneous Dirichlet boundary data. In our problem we choose a function $g(x) \in H^1(0,1)$ such that $g(0) = \alpha$ and such that $g(x)$ is nonzero only on $[0,\xi]$ where $\xi < 1$; the reason for the latter requirement is so that the boundary condition at $x = 1$ is unaffected. We then define $w(x) = u(x) - g(x)$ so that $w(0) = u(0) - g(0) = 0$. Because we have converted the problem to one for $w = u - g$ with $g(x)$ zero outside $[0,\xi]$, $\xi < 1$ we have the same boundary condition for $w'(1)$ as for $u'(1)$. The differential equation is now modified as

$$-\frac{d}{dx}\left(p(x)\frac{d(w+g)}{dx}\right) + q(x)(w+g) = f(x)\,.$$

Because $g(x)$ is a known function, the two-point boundary value problem for $w(x)$

becomes

$$\begin{aligned} -\frac{d}{dx}\left(p(x)\frac{dw}{dx}\right) + q(x)w &= f(x) + (p(x)g'(x))' - q(x)g(x) \quad \text{for } 0 < x < 1 \\ w(0) &= 0 \\ w'(1) + \sigma w(1) &= \beta\,, \end{aligned} \tag{4.43}$$

The mixed boundary condition at the right boundary is handled in a similar manner to the inhomogeneous Neumann. In this case, instead of $w'(1)$ being set to a constant, we have $w'(1) = \beta - \sigma w(1)$. When we substitute this value in the weak form, the constant β goes to the right hand side of the equation because it is known whereas the term $\sigma w(1)$ is unknown and is incorporated in the bilinear form.

We now define a weak problem for the function $w(x) = u(x) - g(x)$. Let $\hat{H}^1(0,1)$ be the subspace of $H^1(0,1)$ consisting of all functions in $H^1(0,1)$ which are zero at $x = 0$. Then we seek a $w \in \hat{H}^1(0,1)$ satisfying

$$\begin{cases} \text{seek } u \in \hat{H}^1(0,1) \text{ satisfying} \\ A(w,v) = (f,v) - A(g,v) + \beta p(1)v(1) \quad \forall\, v \in \hat{H}^1(0,1)\,, \end{cases} \tag{4.44}$$

where

$$A(w,v) = (pw', v') + (qw, v) + \sigma p(1)w(1)v(1)\,. \tag{4.45}$$

To demonstrate that a solution to (4.43) is also a solution to (4.44) we first note that

$$\begin{aligned} \int_0^1 \big[-(pw')' + qw\big]v\,dx &= \int_0^1 pw'v'\,dx + \int_0^1 qwv\,dx - p(1)w'(1)v(1) + p(0)w'(0)v(0) \\ &= \int_0^1 pw'v'\,dx + \int_0^1 qwv\,dx - p(1)\big(\beta - \sigma w(1)\big)v(1) \\ &= A(w,v) - \beta p(1)v(1)\,. \end{aligned}$$

Now the right-hand side of (4.43) can be written as

$$\begin{aligned} &\int_0^1 \big(f(x) + (p(x)g'(x))'\big)v(x)\,dx - \int_0^1 q(x)g(x)v(x)\,dx \\ &\quad = (f,v) - \int_0^1 p(x)g'(x)v'(x)\,dx + p(1)g'(1)v(1) - p(0)g'(0)v(0) \\ &\qquad\quad - \int_0^1 q(x)g(x)v(x)\,dx \\ &\quad = (f,v) - \Big(\int_0^1 p(x)g'(x)v'(x)\,dx + \int_0^1 q(x)g(x)v(x)\,dx\Big) \\ &\quad = (f,v) - A(g,v) \end{aligned}$$

where we have used the fact that $v \in \hat{H}^1(0,1)$ implies $v(0) = 0$ and $g(1) = g'(1) = 0$ because $g = 0$ in $(\xi, 1]$. Combining these two results demonstrates that if w satisfies the classical two-point boundary value problem (4.43) then w satisfies the weak problem (4.44).

Using similar techniques as before, we can demonstrate that $A(\cdot,\cdot)$ defined by (4.45) satisfies the conditions of the Lax-Milgram theorem and that the right-hand side of (4.44) denotes a bounded linear functional on the Hilbert space $\hat{H}^1(0,1)$. Then we have that there exists a unique solution $w \in \hat{H}^1(0,1)$ to (4.44). The generalized or weak solution u to (4.42) is given by $u = w + g$.

To find an approximate solution to (4.44) in the space of piecewise linear functions which are zero at $x = 0$ we define the $(N+1)$-dimensional subspace of $\hat{H}^1(0,1)$

$$\hat{S}^h = \{\phi \in \hat{H}^1(0,1) \,:\, \phi \text{ is piecwise linear on each subinterval and } \phi(0) = 0\}\,,$$

where we are using the mesh defined by

$$0 = x_0 < x_1 < \cdots < x_{N+1} = 1 \quad \text{where} \quad x_i = x_{i-1} + h_i, \quad 1 \le i \le N+1\,,$$

A basis for $\hat{S}^h$ is given by ϕ_i, $i = 1, \ldots, N+1$ where $\phi_1, \ldots, \phi_N$ are defined by (4.8) and ϕ_{N+1} is defined by (4.39). We choose $g(x) = a\phi_0(x)$ where $\phi_0(x)$ is the basis function defined by (4.38). Then the resulting linear system is given by $\mathcal{A}\vec{c} = \vec{b}$ where $\mathcal{A}_{ij} = A(\phi_j, \phi_i)$ for $1 \le i, j \le N+1$ and $\vec{b}_i = -A(a\phi_0, \phi_i) + (f, \phi_i) + bp(1)\phi_i(1)$ for $1 \le i \le N+1$. We note that the coefficient matrix in the resulting algebraic system has the same formulas for the entries as in our previous examples, the dimension is just $N+1$. However, the right-hand side has an additional contribution in the first entry due to the boundary condition at $x = 0$ and in the last position due to the inhomogeneous mixed boundary condition.

We should note that in more complicated problems in higher dimensions, it may not be so easy to construct the function g. To handle the problem theoretically, we can always assume such a function but implementing in a computer program may be more difficult. In later chapters we see different ways to implement inhomogeneous Dirichlet boundary data.

Summarizing, we see that the mixed boundary condition at $x = 1$ required no adjustment of the underlying Hilbert space but rather was "automatically" satisfied by our choice of the weak formulation. As before, such a boundary condition is called *natural*. On the other hand, the Dirichlet boundary condition required that we constrain our underlying Hilbert space so that the boundary condition is satisfied. This is another example of an *essential* boundary condition.

4.4 A fourth order example

In this section we consider approximating the solution of a fourth order boundary value problem. In particular, we consider

$$\begin{aligned} \frac{d^2}{dx^2}\left(r(x)\frac{d^2u}{dx^2}\right) - \frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u(x) &= f(x) \quad 0 < x < 1 \\ u(0) = u(1) = 0 \qquad u''(0) = u''(1) &= 0\,, \end{aligned} \tag{4.46}$$

where $r_{\max} \ge r(x) \ge r_{\min} > 0$ and $p_{\max} \ge p(x) \ge 0$, $q_{\max} \ge q(x) \ge 0$ for all $x \in [0,1]$. Other boundary conditions which can be applied are explored in the exercises.

This problem differs from the previous second order problem because when we perform a a single integration by parts we have three derivatives on the trial function and two on the test function. To balance the derivatives we need to perform a second integration by parts. An obvious choice for the bilinear form $A(\cdot,\cdot)$ is

$$A(v,w) = \int_0^1 \Big(rv''w'' + pv'w' + qvw \Big)\, dx\,. \tag{4.47}$$

In this situation we immediately realize that due to the appearance of second derivatives we can no longer use $H^1(0,1)$ as our underlying Hilbert space; we must now use $H^2(0,1)$ which is the space of all functions in $L^2(0,1)$ which possess weak L^2 derivatives up to order two. The notation $H^2_0(0,1)$ is used for the space

$$H^2_0(0,1) = \{v \in H^2(0,1) \ : \ v(0) = v(1) = v'(0) = v'(1) = 0\}\,. \tag{4.48}$$

Because we have boundary conditions on u'' we don't need $H^2_0(0,1)$ so we consider the space $H^2(0,1) \cap H^1_0(0,1)$ which is the set of all functions v in $H^2(0,1)$ which satisfy $v(0) = 0$ and $v(1) = 0$. Then our weak formulation is to

$$\left\{ \begin{array}{l} \text{seek } u \in H^2(0,1) \cap H^1_0(0,1) \text{ satisfying} \\ \qquad A(u,v) = (f,v) \quad \forall\ v \in H^2(0,1) \cap H^1_0(0,1)\,. \end{array} \right. \tag{4.49}$$

If u is the classical solution of (4.46) then

$$\begin{aligned} (f,v) &= \int_0^1 \Big((ru'')'' - (pu')' + qu \Big) v\, dx \\ &= \int_0^1 \big(-(ru'')'v' + (pu')v' + quv \big)\, dx + ru''v'|_0^1 - pu'v|_0^1 \\ &= \int_0^1 \big(ru''v'' + pu'v' + quv \big)\, dx \\ &= A(u,v)\,, \end{aligned}$$

where we have imposed the boundary conditions $u''(0) = u''(1) = 0$ on the weak form and used the fact that $v \in H^2(0,1) \cap H^1_0(0,1)$ implies $v(0) = v(1) = 0$. In this case the boundary conditions $u''(0) = u''(1) = 0$ are *natural* boundary conditions and $u(0) = u(1) = 0$ are *essential* boundary conditions.

The proof that the bilinear form defined by (4.47) satisfies the hypotheses of the Lax-Milgram theorem is left to the exercises. In the sequel we assume that a unique solution to the weak problem can be guaranteed.

We now consider the approximate problem. An immediate consequence of having $H^2(0,1)$ as the underlying Hilbert space is that we can no longer approximate using continuous piecewise linear polynomials or even continuous piecewise polynomials of degree k. A space S^h consisting of piecewise polynomials satisfies $S^h \subset H^1(0,1)$ if and only if the functions in S^h are continuous; for $S^h \subset H^2(0,1)$ we require the functions and the first derivatives to be continuous. These results are formally proved in a general setting in a later chapter. As a consequence of using $H^2(0,1)$ as the underlying space we must now investigate piecewise polynomials

which are in $C^1(0,1)$ so that we can guarantee them to be subspaces of $H^2(0,1)$. We consider two spaces: piecewise cubic Hermite polynomials and piecewise cubic splines.

4.4.1 Piecewise cubic Hermite polynomials

In this section we consider a space of piecewise polynomials which are $C^1(0,1)$ and which are cubic on each subinterval of a partition of $[0,1]$; for simplicity of exposition we take a uniform partition. We define the space of piecewise cubic Hermite polynomials over the subdivision

$$0 = x_0 < x_1 < \cdots < x_{N+1} = 1 \quad \text{where } x_i = x_{i-1} + h, \quad h = \frac{1}{N+1}$$

to be all polynomials $\phi(x) \in C^1(0,1)$ which are cubic on each subinterval $[x_i, x_{i+1}]$. The dimension of this space is easily determined by considering the number of degrees of freedom and number of constraints we have. On each of the $N+1$ subintervals there are four degrees of freedom to determine a cubic yielding a total of $4N+4$ degrees of freedom; for the piecewise polynomial to be $C^1(0,1)$ we require continuity of the polynomial and its derivative at each of the N interior nodes yielding a total of $2N$ constraints. Combining these results, we see that this space is a $(2N+4)$-dimensional subspace of $H^2(0,1)$. We define the space $\mathcal{H}^h$ to be the space of piecewise cubic Hermite polynomials over the given partition; *i.e.,*

$$\begin{aligned}\mathcal{H}^h = \{\phi(x) \ : \ & \phi \in C^1(0,1), \\ & \phi(x) \text{ is a cubic polynomial on } [x_i, x_{i+1}],\ 0 \le i \le N\}\,. \end{aligned} \tag{4.50}$$

Of course for our particular example we have to constrain this space to satisfy the homogeneous Dirichlet boundary conditions. However, we first consider a basis for $\mathcal{H}^h$.

A convenient way to establish a basis for $\mathcal{H}^h$ is to consider translations of functions defined on $[-1,1]$. In particular, we consider the piecewise cubic polynomials $\xi(x)$ and $\eta(x)$ defined by

$$\xi(x) = \begin{cases} (x+1)^2(-2x+1) & -1 \le x \le 0 \\ (x-1)^2(2x+1) & 0 \le x \le 1 \end{cases}$$

and

$$\eta(x) = \begin{cases} x(x+1)^2 & -1 \le x \le 0 \\ x(x-1)^2 & 0 \le x \le 1 \end{cases}$$

on $[-1,1]$ These polynomials are illustrated in Figure 4.4. Note that $\xi(x) \in C^1[-1,1]$, $\xi(\pm 1) = 0$, $\xi'(0) = 0$, and $\xi'(\pm 1) = 0$; also $\eta(x) \in C^1[-1,1]$, $\eta(0) = \eta(\pm 1) = 0$, $\eta'(\pm 1) = 0$, and $\eta'(0) = 1$.

We now translate these cubic polynomials to the interval $[x_{i-1}, x_{i+1}]$ for $i = 1, \ldots, N$ to obtain our basis elements $\xi_i(x)$ and $\eta_i(x)$. Specifically, we define

$$\xi_i(x) = \begin{cases} \xi(\frac{x}{h} - i) & x_{i-1} \le x \le x_{i+1} \\ 0 & \text{elsewhere} \end{cases} \tag{4.51}$$

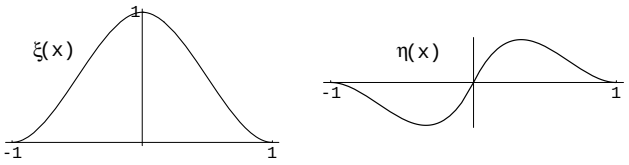

Figure 4.4. *Basis functions for cubic Hermite polynomials on* $[-1, 1]$

and

$$\eta_i(x) = \begin{cases} \eta(\frac{x}{h} - i) & x_{i-1} \le x \le x_{i+1} \\ 0 & \text{elsewhere} \end{cases} \tag{4.52}$$

for $i = 1, \ldots, N$. So far we have $2N$ functions and we know that the dimension of $\mathcal{H}^h$ is $2N + 4$ so we must define four additional functions. To this end, we define $\xi_0(x)$, $\xi_{N+1}(x)$ and $\eta_0(x)$, $\eta_{N+1}(x)$ by

$$\xi_0(x) = \begin{cases} \xi(\frac{x}{h}) & 0 \le x \le x_1 \\ 0 & \text{elsewhere} \end{cases} \qquad \eta_0(x) = \begin{cases} \eta(\frac{x}{h}) & 0 \le x \le x_1 \\ 0 & \text{elsewhere} \end{cases} \tag{4.53}$$

$$\eta_{N+1}(x) = \begin{cases} \eta(\frac{x}{h} - (N+1)) & x_N \le x \le x_{N+1} \\ 0 & \text{elsewhere.} \end{cases} \tag{4.54}$$

and

$$\xi_{N+1}(x) = \begin{cases} \xi(\frac{x}{h} - (N+1)) & x_N \le x \le x_{N+1} \\ 0 & \text{elsewhere} \end{cases} \tag{4.55}$$

Summarizing, we have that

$$\xi_i(x_j) = \delta_{ij} \quad \text{and} \quad \xi_i'(x_j) = 0 \quad \text{for } 0 \le i, j \le N+1 \tag{4.56}$$

and

$$\eta_i(x_j) = 0 \quad \text{and} \quad h\eta_i'(x_j) = \delta_{ij} \quad \text{for } 0 \le i, j \le N+1 \ . \tag{4.57}$$

These polynomials are illustrated in Figure 4.5.

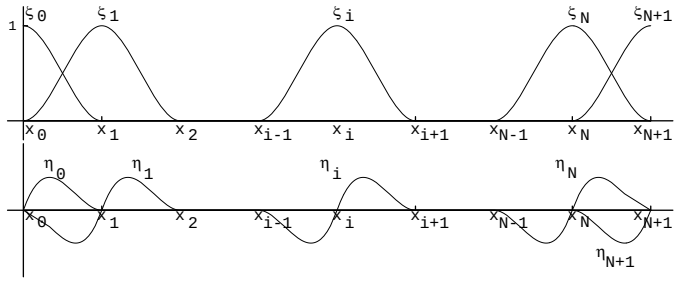

Figure 4.5. *Basis functions for cubic Hermite polynomials*

Clearly these $2N + 4$ functions $\{\xi_i\}_0^{N+1}$, $\{\eta_i\}_0^{N+1}$ belong to $\mathcal{H}^h$; moreover, they form a basis for $\mathcal{H}^h$. To see this, let $p(x) \in \mathcal{H}^h$ so that $p \in C^1(0, 1)$, $p(x)$ is a

cubic on each subinterval $[x_i, x_{i+1}]$, $0 \le i \le N$. Clearly $p(x)$ is uniquely determined by its value and that of its derivative at the $N+2$ nodes $x_0, \ldots, x_{N+1}$. Using (4.56) and (4.57) we have

$$p(x) = \sum_{i=0}^{N+1} p(x_i)\xi_i(x) + h \sum_{i=0}^{N+1} p'(x_i)\eta_i(x)\,.$$

Thus the vectors span $\mathcal{H}^h$ and are also clearly linearly independent.

Of course, for our example, we must constrain the space to satisfy the homogeneous Dirichlet boundary conditions. To this end, we define $\widehat{\mathcal{H}}^h$ to be all functions $\phi(x) \in \mathcal{H}^h$ which satisfy $\phi(0) = \phi(1) = 0$. In this case we choose the $2n+2$ functions $\{\xi_i\}_1^N$, $\{\eta_i\}_0^{N+1}$; we do not include ξ_0, ξ_{N+1} since from (4.56) we know that $\xi_0(0) = 1$, $\xi_{N+1}(1) = 1$ so that $\xi_0, \xi_{N+1} \notin \widehat{\mathcal{H}}^h$.

We can now pose our weak problem over $\widehat{\mathcal{H}}^h \subset H^2(0,1) \cap H_0^1(0,1)$. We seek $u^h \in \widehat{\mathcal{H}}^h$ satisfying

$$A(u^h, v^h) = \int_0^1 \left(r u^{h\prime\prime} v^{h\prime\prime} + p u^{h\prime} v^{h\prime} + q u^h v^h \right) dx = \left(f, v^h\right) \quad \forall\, v^h \in \widehat{\mathcal{H}}^h\,. \quad (4.58)$$

Once we have chosen a basis for our approximating space, we know that our discrete weak problem reduces to solving a linear system of algebraic equations $\mathcal{A}c = \mathcal{F}$. We let $\{\phi_i\}_{i=1}^{2N+2}$ be the basis functions $\{\xi_i(x), \eta_i(x)\}$ defined by (4.51)–(4.54) and ordered in the sequence $\{\eta_0, \xi_1, \eta_1, \xi_2, \eta_2, \ldots, \xi_N, \eta_N, \eta_{N+1}\}$. If we write $u^h = \sum_{i=1}^{2N+2} c_j \phi_j(x)$ then the c_j's represent either the nodal values of u^h or of $h(u^h)'$. The matrix $\mathcal{A}$ whose entries are given by

$$\mathcal{A}_{ij} = \int_0^1 \left(r(x)\phi_i''(x)\phi_j''(x) + p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x) \right) dx$$

is a symmetric matrix. However, the matrix is no longer tridiagonal as in the case of piecewise linear elements but rather has the block tridiagonal form

$$\begin{pmatrix} A_0 & B_0 & 0 & & \cdots & 0 \\ B_0 & A_1 & B_1 & 0 & \cdots & 0 \\ 0 & B_1 & A_2 & B_2 & 0 & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & B_{N-2} & A_{N-1} & B_{N-1} \\ 0 & \cdots & & 0 & B_{N-1} & A_N \end{pmatrix}$$

where the A_i's and B_i's are 2×2 matrices. To see this, consider the interval $[x_{i-1}, x_{i+1}]$. The basis functions which are nonzero on this interval are ξ_{i-1}, η_{i-1}, ξ_i, η_i, ξ_{i+1}, and η_{i+1} so that the maximum number of nonzero entries in a single row is six. It can be shown that the coefficient matrix is also positive definite so that the linear system can be efficiently solved using a block Cholesky factorization.

In order to obtain an error estimate we turn to Galerkin's theorem which provides us with the H^2-estimate

$$\left\| u - u^h \right\|_2 \le \inf_{\chi^h \in \widehat{\mathcal{H}}^h} \left\| u - \chi^h \right\|_2\,, \quad (4.59)$$

where u and u^h satisfy (4.49) and (4.58), repectively. To bound the term for the error in the best approximation, we consider the $\mathcal{H}^h$ interpolant. The Hermite cubic interpolant of a function $g(x)$ on the uniform partition of $[0,1]$

$$0 = x_0 < x_1 < \cdots < x_{N+1} = 1 \quad \text{where} \quad x_i = x_{i-1} + h_i, \quad 1 \le i \le N+1\,,$$

where $g(0) = g(1) = 0$ is given by

$$I^h g = \sum_{i=1}^{N} g(x_i)\xi_i(x) + h \sum_{i=0}^{N+1} g'(x_i)\eta_i(x)\,. \tag{4.60}$$

From approximation theory we have the following result which tells us how well a function can be approximated by its cubic Hermite interpolant. As in the case of the piecewise linear interpolant, additional smoothness on the function must be assumed in order to get the optimal rates of approximation.

Lemma 4.12. *Let $f \in H^s(0,1)$ where $2 \le s \le 4$ and let $I^h f$ denote its piecewise cubic Hermite interpolant defined by (4.60). Then for $0 \le k \le 2$ and for some constant C, we have that*

$$\left\| f - I^h f \right\|_k \le C h^{s-k} \left\| f \right\|_s . \tag{4.61}$$

Note that in order to get the optimal accuracy, *e.g.,* $O(h^4)$ in the L^2-norm, we must have that $f \in H^4(0,1)$.

We can now use Lemma 4.12 to bound the right-hand side of (4.59). We have the following error estimates; the proof is similar to that of Theorems 4.3 and 4.4.

Theorem 4.13. *Let $u \in H^k(0,1)$, $2 \le k \le 4$, be the solution of (4.49) and let $u^h \in \widehat{\mathcal{H}}^h$ be the solution of (4.58). Then*

$$\left\| u - u^h \right\|_j \le C h^{k-j} \left\| u \right\|_k \tag{4.62}$$

where $j = 0$ or $j = 2$.

In practice, cubic Hermite functions are not often used. The reason for this is twofold. First, there are, in general, $2N + 4$ parameters to compute in one dimension; as we will see in the next section there is an approximating space which maintains the same accuracy as cubic Hermites with only $N+4$ parameters. Hence for cubic Hermite polynomials we would be solving a $(2N+4)$-dimensional system in $\mathbb{R}^1$ versus a $(N+4)$-dimensional system; this difference in size of the linear system magnifies as we move to higher dimensions. Secondly, the Hermite cubic interpolant matches the function and its derivative at the nodes. In many situations, especially in more than one-dimension, it is not possible to accurately specify the derivatives at the nodes. The space considered in the next section requires interpolation of the function values only.

4.4.2 Piecewise cubic spline functions

In this section we consider a finite dimensional subspace of $H^2(0,1)$ which has two desirable properties. We will require that only function values (and not derivatives) will be used to interpolate smooth functions and that our space has as small a dimension as possible. To do this, we constrain the space of Hermite cubics to obtain the space

$$\mathcal{C}^h = \{\phi(x) \ : \ \phi(x) \in C^2[0,1], \ \phi(x)\text{is a cubic in each } [x_i, x_{i+1}]\}\,, \tag{4.63}$$

where we are using the same uniform partition defined by $h = 1/(N+1)$ as before. In an analogous manner to the case of $\mathcal{H}^h$ of cubic Hermite polynomials, we can determine that $\mathcal{C}^h$ is a $(N+4)$-dimensional subspace of $H^2(0,1)$.

We must now specify a basis for $\mathcal{C}^h$. In our space $\mathcal{H}^h$ of cubic Hermite polynomials, there was a clear criterion for determining its elements. In fact, to determine $\phi \in \mathcal{H}^h$ we just specified $\phi(x_i)$ and $\phi_i'(x_i)$ at the nodes x_i, $0 \le i \le N+1$, thus defining a unique cubic polynomial on each interval and at the same time assuring that it be in $C^1[0,1]$. For the cubic spline space $\mathcal{C}^h$, the obvious thing to try in order to assure C^2-continuity at the nodes would be to specify $\phi(x)$, $\phi'(x)$, and $\phi''(x)$ there. However, this cannot be done using cubic polynomials because we would be overspecifying them.

Instead of specifying basis functions which interpolate a function and its first and second derivatives at the nodes, we take the approach of constructing a basis for $\mathcal{C}^h$. To find the ith basis function we first note that its support cannot be in the interval $[x_{i-1}, x_{i+1}]$ as was the case with piecewise linear functions and Hermite cubic functions. To see this, we note that there are eight degrees of freedom to determine the cubic polynomials on $[x_{i-1}, x_{i+1}]$ and a total of nine conditions to specify over the three nodes, *i.e.,* $\phi(x_{i\pm1}) = \phi'(x_{i\pm1}) = \phi''(x_{i\pm1}) = 0$ and the continuity of $\phi(x)$, $\phi'(x)$, and $\phi''(x)$ at $x = x_i$. Consequently, we must extend our interval to $[x_{i-2}, x_{i+2}]$ and attempt to construct a C^2 function which is cubic on each of the four subintervals $[x_{i-s}, x_{i-s+1}]$, for $s = -1, 0, 1, 2$ and which is zero outside the interval $[x_{i-2}, x_{i+2}]$. In this case we have 16 degrees of freedom and 15 conditions to impose so that it is clearly possible; the extra degree of freedom will be used to specify that the function is one at node x_i. A straightforward, but tedious, computation gives such a function on the interval $[-2, 2]$; this function is illustrated in Figure 4.6. Translating this function to the interval $[x_{i-2}, x_{i+2}]$ for $2 \le i \le N-1$ we have

$$\phi_i(x) = \begin{cases} \phi(\frac{x}{h} - i) & x_{i-2} \le x \le x_{i+2} \\ 0 & \text{elsewhere} \end{cases} \tag{4.64}$$

where

$$\phi(x) = \begin{cases} \frac{1}{4}(x+2)^3 & -2 \le x \le -1\,, \\ \frac{1}{4}\left(1+3(1+x)+3(1+x)^2-3(1+x)^3\right) & -1 \le x \le 0\,, \\ \frac{1}{4}\left(1+3(1-x)+3(1-x)^2-3(1-x)^3\right) & 0 \le x \le 1\,, \\ \frac{1}{4}(2-x)^3 & 1 \le x \le 2\,. \end{cases} \tag{4.65}$$

Clearly each $\phi_i(x) \in \mathcal{C}^h$ for $2 \le i \le N-1$; we have a total of $N-2$ functions so

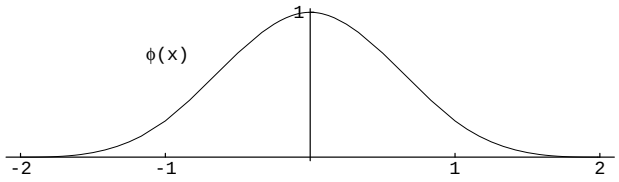

Figure 4.6. *Basis function on* $[-2,2]$ *for cubic splines*

that an additional six basis functions are needed to reach the dimension $N+4$. We add the functions defined below where we have introduced extra nodes $x_{-1} = -h$ and $x_{N+2} = 1+h$:

$$\phi_{-1}(x) = \begin{cases} \phi(\frac{x}{h}+1) & 0 \le x \le x_1 \\ 0 & \text{otherwise} \end{cases} \qquad \phi_0(x) = \begin{cases} \phi(\frac{x}{h}+1) & 0 \le x \le x_2 \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_1(x) = \begin{cases} \phi(\frac{x}{h}-1) & 0 \le x \le x_3 \\ 0 & \text{otherwise} \end{cases} \qquad \phi_N(x) = \begin{cases} \phi(\frac{x}{h}-N) & x_{N-2} \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{N+1}(x) = \begin{cases} \phi(\frac{x}{h}-N-1) & x_{N-1} \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.66}$$

$$\phi_{N+2}(x) = \begin{cases} \phi(\frac{x}{h}-N-2) & x_N \le x \le 1 \\ 0 & \text{otherwise}\,. \end{cases}$$

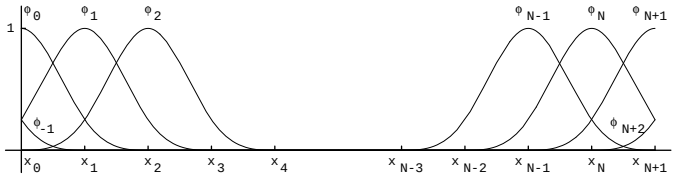

Figure 4.7. *Basis functions for cubic splines*

The set $\{\phi_i(x)\}_{i=-1}^{N+2}$ form a basis for $\mathcal{C}^h$. These basis functions are illustrated in Figure 4.7.

An important difference in the basis functions for cubic splines and those we studied for piecewise linear functions and cubic Hermites is that the unknowns we solve for are no longer the nodal values of u^h. This is because the cubic spline basis functions are no longer zero at all nodes except one.

Since for our problem the underlying Hilbert space is $H^2(0,1) \cap H_0^1(0,1)$, *i.e.,* all functions in $H^2(0,1)$ which are zero at $x = 0$ and at $x = 1$, we need only $N + 2$ basis functions. We define $\widehat{\mathcal{C}}^h$ as

$$\widehat{\mathcal{C}}^h = \{\phi \in \mathcal{C}^h \ : \ \phi(0) = \phi(1) = 0\} .$$

For cubic splines we can not simply omit the specific basis functions which are nonzero at $x = 0$ and $x = 1$ as we did for the piecewise linear basis. We can equip $\widehat{\mathcal{C}}^h$ with a basis consisting of $\{\phi_i\}_{i=2}^{N-1}$ and the four functions $\tilde{\phi}_0$, $\tilde{\phi}_1$, $\tilde{\phi}_N$, and $\tilde{\phi}_{N+1}$ which are obtained as linear combinations of the remaining $\phi_i(x)$'s so that they vanish at $x = 0$ and at $x = 1$. For example, $\tilde{\phi}_0 = \phi_0 - 4\phi_{-1}$, $\tilde{\phi}_1 = \phi_1 - \phi_{-1}$. (See exercises.)

As before, in order to obtain error estimates using cubic splines as our approximating space, we need to obtain estimates for the error in the cubic spline interpolant. We note that we have $N + 2$ nodes and the space $\mathcal{C}^h$ has dimension $N + 4$; thus if the interpolant matches the function value at the $N + 2$ nodes, then we have two additional conditions to impose. There are numerous choices; here we consider one type of cubic spline interpolant of a function $f \in H^2(0,1)$. We require the interpolant, denoted $I^h f$, to satisfy the $N + 4$ conditions

$$\begin{array}{rcll} I^h f(x_i) & = & f(x_i) & \text{for } 0 \le i \le N+1 \\ I^h f'(x_0) & = & f'(x_0) , & I^h f'(x_{N+1}) = f'(x_{N+1}) . \end{array} \tag{4.67}$$

Lemma 4.14. *Let $f \in H^s(0,1)$, $2 \le s \le 4$. Then for $0 \le r \le 2$ we have that*

$$\left\| D^r (f - I^h f) \right\|_0 \le C h^{s-r} \left\| D^s f \right\| . \tag{4.68}$$

We now state a result analogous to Theorem 4.13 .

Theorem 4.15. *Let $u \in H^k(0,1)$, $2 \le k \le 4$ be the solution of (4.49) and let $u^h \in \widehat{\mathcal{C}}^h$ be the solution of (4.58). Then*

$$\left\| u - u^h \right\|_j \le C h^{k-j} \left\| u \right\|_k \tag{4.69}$$

where $j = 0$ or $j = 2$.