Chapter

4 Variational approximation methods for elliptic PDEs

One of the virtues of the variational approach is that it leads naturally to a whole family of approximation methods. The reason why approximation methods for PDEs are needed is that, even though we may be able to prove the existence of a solution, in general there is no closed form formula for it. Therefore, if we need quantitative information on the solution, there is little choice but to try to approximate it with something that is computable in practice. Let us point out that, even though we limit ourselves to variational approximation methods, there are other approximation methods that are not variational.

4.1 The general abstract variational approximation scheme

As we have seen, boundary value problems take place in infinite dimensional vector spaces. An infinite dimensional space is way too big to fit inside a computer, thus the main idea is to build finite dimensional approximations. Any approximation method of this kind falls under the general heading of a *Galerkin method*. Let us start with a few definitions that pertain to the variational case.

Definition 4.1.1 *Let V be a Hilbert space and $(V_n)_{n\in\mathbb{N}}$ be a sequence of finite dimensional vector subspaces of V. We say that this sequence is a* conforming approximation sequence *if for all $u \in V$, there exists a sequence $(v_n)_{n\in\mathbb{N}}$ such that*

$$v_n \in V_n \text{ and } \|u - v_n\|_V \to 0 \text{ when } n \to +\infty. \tag{4.1}$$

Remark 4.1.1 Note that in general, we do not have $V_n \subset V_{n+1}$, *i.e.*, the spaces do not need to be nested. The conforming approximation condition implies that $\bigcup_{n\in\mathbb{N}} V_n$ is dense in V.

There are situations in which *non conforming approximations* are called for, that is to say $V_n \not\subset V$. Of course, in this case $\|u - v_n\|_V$ does not make sense, and another definition is needed.

The traditional notation for an approximation sequence is V_h where h is a discretization parameter that is assumed to belong to a sequence that tends to 0. We will from now on stick with the tradition. □

The main abstract result is the following, also known under the name of Céa's lemma.

Theorem 4.1.1 *Let V be a Hilbert space, a be a bilinear form and ℓ be a linear form satisfying the hypotheses of the Lax-Milgram theorem. Let V_h be a closed subspace of V. Then there exists a unique $u_h \in V_h$ such that*

$$\forall v_h \in V_h, \quad a(u_h, v_h) = \ell(v_h),$$

and we have

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \inf_{v_h\in V_h} \|u - v_h\|_V = \frac{M}{\alpha} d(u, V_h),$$

where M is the continuity constant of a and α its V-ellipticity constant.

Proof. Since V_h is closed, it is a Hilbert space for the restriction of the scalar product of V. The Lax-Milgram hypotheses for the variational problem on V_h are thus satisfied and the existence and uniqueness of u_h is assured.

Now we have $a(u, v) = \ell(v)$ for all $v \in V$, thus in particular for $v = w_h \in V_h$. On the other hand, we also have $a(u_h, w_h) = \ell(w_h)$, so that subtracting the two

$$0 = a(u, w_h) - a(u_h, w_h) = a(u - u_h, w_h)$$

for all $w_h \in V_h$. By V-ellipticity, for all $v_h \in V_h$,

$$\begin{aligned}
\alpha\|u - u_h\|_V^2 &\leq a(u - u_h, u - u_h) \\
&\leq a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\
&= a(u - u_h, u - v_h) \\
&\leq M\|u - u_h\|_V \|u - v_h\|_V,
\end{aligned}$$

since $v_h - u_h \in V_h$. The case $\|u - u_h\|_V = 0$ is ideal and nothing needs to be done. If it is non zero, we divide by it and obtain

$$\|u - u_h\|_V \leq \frac{M}{\alpha}\|u - v_h\|_V$$

for all $v_h \in V_h$, thus the theorem by taking the infimum of the right-hand side. □

Corollary 4.1.2 *Let V_h be a conforming approximation sequence. Then the sequence $u_h \in V_h$ of approximated solutions converges to the solution u in V, with the* a priori error estimate

$$\|u - u_h\|_V \leq \frac{M}{\alpha} d(u, V_h) \to 0 \textit{ when } h \to 0.$$

Proof. Each subspace V_h is finite dimensional, hence closed. We thus apply Theorem 4.1.1 and obtain the convergence result since $d(u, V_h) \leq \|u - v_h\|_V$ where v_h is given by the definition of conforming approximation for this u. □

Remark 4.1.2 We also trivially have $\|u - u_h\|_V \geq d(u, V_h)$, thus the error estimate is optimal in terms of order of magnitude when $h \to 0$. Now, if the constant M/α is very large, then the numerical error can be large too with respect to $d(u, V_h)$.

An interesting feature of Céa's lemma is that it decomposes the error estimate into two basically independent parts: The constant M/α which only depends on the bilinear form, *i.e.*, the PDE, and not on the approximation method, and $d(u, V_h)$ which depends mostly on the approximation properties of the space V_h. In practice, the second part will be estimated by constructing a linear operator $\Pi_h \colon V \to V_h$, writing that

$$d(u, V_h) \leq \|u - \Pi_h u\|_V \leq \|I - \Pi_h\| \|u\|_V$$

and estimating the term $\|I - \Pi_h\|$ which depends only on V_h. □

The approximation u_h lives in a finite dimensional space, therefore it is computable, at least in principle. Let us see how to proceed in practice.

Proposition 4.1.1 *Let $N_h = \dim V_h$ and $(w^1, w^2, \dots, w^{N_h})$ be a basis of V_h. We write $u_h = \sum_{j=1}^{N_h} u_{h,j} w^j$. We introduce an $N_h \times N_h$ matrix A defined by $A_{ij} = a(w^j, w^i)$ and two vectors $B \in \mathbb{R}^{N_h}$ by $B_i = \ell(w^i)$ and $X \in \mathbb{R}^{N_h}$ by $X_j = u_{h,j}$. Then the matrix A is invertible and we have $AX = B$. Conversely, the solution of this linear system is the vector of coordinates of u_h in the basis (w^i).*

Proof. Let us take $v_h = w^i$ in the variational formulation of the finite dimensional problem. This yields

$$B_i = \ell(w^i) = a(u_h, w^i) = a\Big(\sum_{j=1}^{N_h} u_{h,j} w^j, w^i\Big) = \sum_{j=1}^{N_h} u_{h,j} a(w^j, w^i) = \sum_{j=1}^{N_h} A_{ij} X_j = (AX)_i$$

for all i. Hence $AX = B$.

Conversely, if $AX = B$, then by the above computation, $\ell(w^i) = a(\widetilde{u}_h, w^i)$ where $\widetilde{u}_h = \sum_{j=1}^{N_h} X_j w^j$. For all $v_h \in V_h$, we have $v_h = \sum_{i=1}^{N_h} v_{h,i} w^i$, therefore

$$\ell(v_h) = \sum_{i=1}^{N_h} v_{h,i} \ell(w^i) = \sum_{i=1}^{N_h} v_{h,i} a(\widetilde{u}_h, w^i) = a\Big(\widetilde{u}_h, \sum_{i=1}^{N_h} v_{h,i} w^i\Big) = a(\widetilde{u}_h, v_h)$$

therefore, by the uniqueness of the Lax-Milgram solution, we have $\widetilde{u}_h = u_h$. Thus the variational problem and the linear system are equivalent. Since the variational problem has one and only one solution, it follows that A is invertible. □

Remark 4.1.3 The problem of computing the finite dimensional approximation u_h is thus reduced to that of computing the matrix A and the right-hand side B once a basis of V_h is chosen, which is called *assembling the system*, and then of solving the linear system $AX = B$. In practical applications, N_h is typically large, ranging from the thousands to the millions. This is a whole other subject with many facets: matrix conditioning, efficient algorithms for large linear systems, high performance computing. We will not touch on this.

It is important not to loose sight of the fact that the size of the matrix A and of the right-hand side B depend on h, via N_h, even though the notation fails to make this dependence apparent. In particular, when $h \to 0$, we have $N_h \to +\infty$.

Do not forget the exchange of indices $A_{ij} = a(w^j, w^i)$ and *not* $A_{ij} = a(w^i, w^j)$! Note that if a is symmetric, then the matrix A is symmetric, positive, definite, with $a(v, v) = Y^T AY$ where Y is the vector of coordinates of v in the basis (w^i). □

We now introduce the main example of variational approximation method, the finite element method (FEM). For simplicity, we start with the one-dimensional case.

4.2 The finite element method in dimension one

Let $\Omega =]a, b[$ and consider the model problem

$$\begin{cases} -u'' + cu = f \text{ in } \Omega, \\ u(a) = u(b) = 0. \end{cases} \tag{4.2}$$

When $f \in L^2(a,b)$, $c \in L^\infty(a,b)$ and $c \geq 0$, we know that this problem has one and only one solution by using the variational formulation $V = H_0^1(]a,b[)$, $a(u,v) = \int_\Omega (u'v' + cuv)\,dx$ and $\ell(v) = \int_\Omega fv\,dx$.

The idea of the FEM is to take approximation spaces V_h composed of functions that are piecewise polynomial of low degree, with lots of pieces. In one dimension, we have $H^1 \subset C^0$, thus for the approximation to be conforming, we need to impose $V_h \subset C^0$ as well.

The FEM is based on the notion of *mesh*. In one dimension, a mesh is just a subdivision of $]a,b[$ into a finite number of intervals. Each of the small intervals is called an *element*. We will only consider uniform meshes. Let N be an integer. We set $h = \frac{b-a}{N+1}$, which is called the *mesh size*, and let $x_i = a + ih$, $i = 0,\dots,N+1$, be the *nodes* of the mesh.

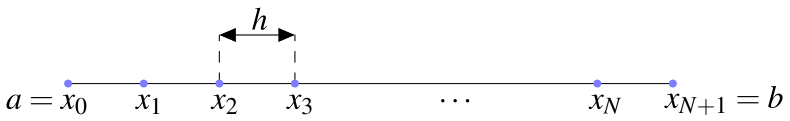

Figure 1. A uniform 1d mesh.

We thus have $N+1$ subintervals $[x_i, x_{i+1}]$ of length h, N interior nodes x_i, $i = 1,\dots,N$, and 2 boundary nodes x_0 and x_{N+1}. We now define

$$V_h = \{v_h \in C^0([a,b]); v_{h|[x_i,x_{i+1}]} \text{ is affine for } i = 0,\dots,N, v_h(a) = v_h(b) = 0\}. \tag{4.3}$$

Note that here, the subscript h in V_h is actually the same h as the mesh size. Since $h \to 0$ when $N \to +\infty$, we thus have a sequence of approximation spaces. We first need to see if this would-be approximation is conforming.

Proposition 4.2.1 *We have $V_h \subset H_0^1(]a,b[)$.*

Proof. First of all, since $V_h \subset C^0([a,b])$ with $[a,b]$ compact, we have $V_h \subset L^2(a,b)$. Let us compute the distributional derivative of an element v_h of V_h. Since $v_{h|[x_i,x_{i+1}]}$ is an affine function, we can write $v_h(x) = \lambda_i x + \mu_i$ for $x \in [x_i, x_{i+1}]$, with λ_i, μ_i constants that depend on the subinterval. For all $\varphi \in \mathscr{D}(]a,b[)$, we have

$$\begin{aligned}\langle v_h', \varphi\rangle = -\langle v_h, \varphi'\rangle &= -\int_a^b v_h(x)\varphi'(x)\,dx \\ &= -\sum_{i=0}^N \int_{x_i}^{x_{i+1}} v_h(x)\varphi'(x)\,dx \\ &= -\sum_{i=0}^N \int_{x_i}^{x_{i+1}} (\lambda_i x + \mu_i)\varphi'(x)\,dx.\end{aligned}$$

Now we can classically integrate each element integral by parts,

$$\begin{aligned}-\int_{x_i}^{x_{i+1}}(\lambda_i x+\mu_i)\varphi'(x)\,dx &= \int_{x_i}^{x_{i+1}}\lambda_i\varphi(x)\,dx-[v_h(x)\varphi(x)]_{x_i}^{x_{i+1}}\\ &= \int_{x_i}^{x_{i+1}}\lambda_i\varphi(x)\,dx-v_h(x_{i+1})\varphi(x_{i+1})+v_h(x_i)\varphi(x_i),\end{aligned}$$

since v_h is continuous on $[a,b]$ its right and left limits at x_i and x_{i+1} respectively are just its value at these points.

Now, if we let

$$g(x)=\sum_{i=0}^{N}\lambda_i\mathbf{1}_{[x_i,x_{i+1}]}(x),$$

then obviously g is a piecewise constant function hence is bounded, and thus in $L^2(a,b)$ and

$$\sum_{i=0}^{N}\int_{x_i}^{x_{i+1}}\lambda_i\varphi(x)\,dx=\int_a^b g(x)\varphi(x)\,dx.$$

On the other hand

$$\begin{aligned}-\sum_{i=0}^{N}[v_h(x)\varphi(x)]_{x_i}^{x_{i+1}} &= -v_h(x_1)\varphi(x_1)+v_h(x_0)\varphi(x_0)\\ &\quad -v_h(x_2)\varphi(x_2)+v_h(x_1)\varphi(x_1)-\cdots\\ &\quad \cdots-v_h(x_{N+1})\varphi(x_{N+1})+v_h(x_N)\varphi(x_N)=0,\end{aligned}$$

since all terms involving interior nodes appear twice with opposite signs, and $\varphi(x_0)=\varphi(x_{N+1})=0$ since φ has compact support. Finally, we see that

$$\langle v_h',\varphi\rangle=\int_a^b g(x)\varphi(x)\,dx=\langle g,\varphi\rangle,$$

with $g\in L^2(]a,b[)$ which shows that $v_h\in H^1(]a,b[)$ and $v_h'=g$. Now all elements of V_h also satisfy $v_h(a)=v_h(b)=0$ so that $v_h\in H_0^1(]a,b[)$. □

It is fairly clear that the space V_h is finite dimensional, since any of its elements is determined by a finite number of constants λ_i and μ_i. Therefore, the general abstract principle applies and there exists a unique $u_h\in V_h$ such that $a(u_h,v_h)=\ell(v_h)$ for all $v_h\in V_h$, and we have Céa's lemma error estimate. Let us see how this estimate can be exploited to quantify the convergence rate. Let us start with a general purpose lemma concerning V_h.

Lemma 4.2.1 *There exists a unique continuous linear mapping* $\Pi_h\colon H_0^1(]a,b[)\to V_h$*, called the* V_h-interpolation operator *such that for all* v *in* $H_0^1(]a,b[)$*,* $v(x_i)=\Pi_h v(x_i)$ *for* $i=0,\dots,N+1$.

Proof. First of all, we note that $H_0^1(]a,b[) \hookrightarrow C^0([a,b])$, therefore the nodal values $v(x_i)$ are unambiguously defined and $v(x_0) = v(x_{N+1}) = 0$.

Now an affine function on $[x_i, x_{i+1}]$ is uniquely determined by its values at x_i and x_{i+1}. Thus, the relations $v(x_i) = \Pi_h v(x_i)$ for $i = 0, \dots, N+1$ define a unique piecewise affine function on the mesh, that is continuous and vanishes at both ends, thus belongs to V_h. Let $\Pi_h v$ be this function. Clearly, the mapping $v \mapsto \Pi_h v$ is linear from $H_0^1(]a,b[)$ into V_h. Finally we infer from the fact that the values taken by an affine function on an interval lie between the values at the endpoints, that $\max_{x\in[x_i,x_{i+1}]} |\Pi_h v(x)| = \max\big(|v(x_i)|, |v(x_{i+1})|\big)$, and therefore

$$\begin{aligned}\|\Pi_h v\|_{C^0([a,b])} &= \max_{i=0,\dots,N} \max_{x\in[x_i,x_{i+1}]} |\Pi_h v(x)| \\ &= \max_{i=0,\dots,N} \max\big(|v(x_i)|, |v(x_{i+1})|\big) \\ &\le \max_{x\in[a,b]} |v(x)| = \|v\|_{C^0([a,b])} \le C\|v\|_{H^1(]a,b[)},\end{aligned}$$

by Theorem 2.7.1. Consequently, the V_h-interpolation operator is continuous. □

Remark 4.2.1 A picture is in order here.

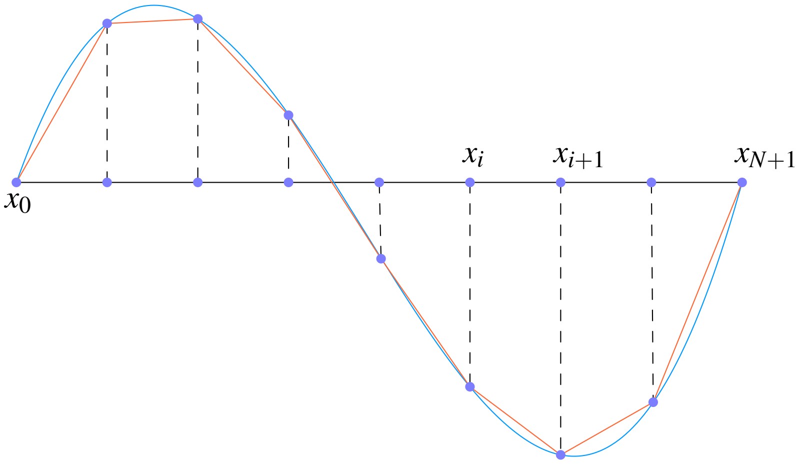

Figure 2. The V_h-interpolate $\Pi_h v$ of a function v.

In other words, $\Pi_h v$, which we call the *V_h-interpolate of v*, is the unique element of V_h that coincides with v at all nodes of the mesh. □

Proposition 4.2.2 *Assume c and f are continuous on $[a,b]$. Then u is of class $C^2([a,b])$ and there exists a constant C independent of u such that*

$$\|u - u_h\|_V \le Ch \max_{[a,b]} |u''|. \tag{4.4}$$

Proof. If f and c are continuous, since u is also continuous by Theorem 2.7.1, then $u'' = cu - f$ is continuous on $[a,b]$ and $u \in C^2([a,b])$. By Céa's lemma (aka Theorem 4.1.1), we have

$$\|u-u_h\|_V \leq \frac{M}{\alpha} \inf_{v_h\in V_h} \|u-v_h\|_V.$$

We choose $v_h = \Pi_h u$. It follows that

$$\|u-u_h\|_V \leq \frac{M}{\alpha}\|u-\Pi_h u\|_V,$$

and we are left with estimating the rightmost norm.

Let us take the H^1 semi-norm as a norm on V (this makes for simpler computations). We have

$$\|u-\Pi_h u\|_V^2 = \int_a^b ((u-\Pi_h u)')^2\,dx = \sum_{i=0}^N \int_{x_i}^{x_{i+1}} (u'-(\Pi_h u)')^2\,dx.$$

Let us consider the function $w = u - \Pi_h u$ on $[x_i, x_{i+1}]$. By definition of V_h-interpolation, we have $w(x_i) = w(x_{i+1}) = 0$. Since w is C^1 on $[x_i, x_{i+1}]$, Rolle's theorem applies and there exists $c \in]x_i, x_{i+1}[$ such that $w'(c) = 0$. Now, w is also of class C^2 on $[x_i, x_{i+1}]$ so that

$$w'(x) = \int_c^x w''(t)\,dt = \int_c^x u''(t)\,dt$$

for all $x \in [x_i, x_{i+1}]$, since $\Pi_h u$ is affine there, thus its second derivative vanishes. It follows from this equality that

$$|w'(x)| \leq \int_c^x |u''(t)|\,dt \leq \int_{x_i}^{x_{i+1}} |u''(t)|\,dt \leq h \max_{t\in[x_i,x_{i+1}]} |u''(t)| \leq h \max_{[a,b]} |u''|,$$

for all $x \in [x_i, x_{i+1}]$. Squaring and integrating, we thus see that

$$\int_{x_i}^{x_{i+1}} (u'-(\Pi_h u)')^2\,dx = \int_{x_i}^{x_{i+1}} (w')^2\,dx \leq h^3 \max_{[a,b]} |u''|^2.$$

Now we sum from $i = 0$ to N

$$\|u-\Pi_h u\|_V^2 \leq h^3 \Big(\sum_{i=0}^N 1\Big) \max_{[a,b]} |u''|^2 = h^3(N+1) \max_{[a,b]} |u''|^2.$$

At this point, we recall that $h = \frac{b-a}{N+1}$, hence

$$\|u-\Pi_h u\|_V^2 \leq h^2(b-a) \max_{[a,b]} |u''|^2,$$

and finally

$$\|u-u_h\|_V \leq \Big(\frac{M}{\alpha}\sqrt{b-a}\Big) h \max_{[a,b]} |u''|,$$

which completes the proof. □

Remark 4.2.2 Note that we have not proved that the sequence V_h is a conforming approximation sequence in the sense of Definition 4.1.1. Rather, we have exploited Céa's error estimate directly, coupled with an additional regularity hypothesis, here that u be C^2 essentially, to obtain an explicit error estimate and a convergence order in $O(h)$ when $h \to 0$. The sequence V_h is in fact a conforming approximation sequence, but this does not turn out to be too useful, as the convergence toward a generic element of H^1 can be much slower than $O(h)$. This will be a general fact: additional regularity hypotheses on the solution will be needed for explicit error estimates. Such regularity can however often be deduced from elliptic regularity theory. □

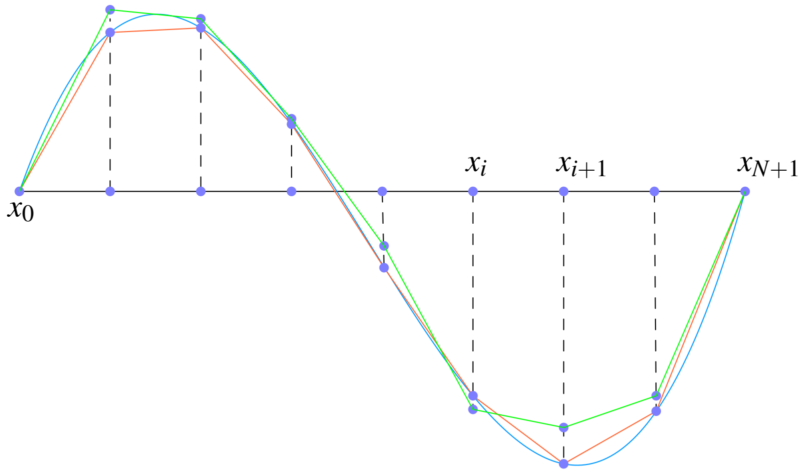

Figure 3. A fictitious computation : the continuous solution u in blue, the discrete solution u_h in green, and the V_h-interpolate $\Pi_h u$ of u in orange used to control the error between the former two. Note that only u_h is effectively computable.

Let us now talk about the choice of a basis in V_h. Even though in principle, the resolution of the finite dimensional problem should not depend on the basis choice, in practice this is an extremely important issue since the choice of basis directly impacts the matrix A. A wrong choice of basis can lead to a linear problem that cannot be solved numerically (bad conditioning, full matrix) in the sense that all theoretically convergent algorithms may fail or take too long or use up too much computer memory. Recall that for a basis (w^j), the matrix coefficients are given by

$$A_{ij} = a(w^j, w^i) = \int_a^b ((w^j)'(w^i)' + cw^j w^i)\,dx.$$

For numerical purposes, full matrices are to be avoided and sparse matrices preferred. A sparse matrix is a matrix in which most coefficients are 0 and nonzero coefficients are close to the diagonal. Now, there is an easy way of making sure that $A_{ij} = 0$, given the above formula, and that is to arrange that the supports of w^i and w^j have negligible intersection. So we want to find a basis for V_h for which the supports are as small as possible, in order to minimize the intersections. Now clearly, the support of any function of V_h is at least comprised of two elements. We thus define

Definition 4.2.1 *For $i = 1,\dots,N$, let $w_h^i \in V_h$ be defined by $w_h^i(x_i) = 1$ and $w_h^i(x_j) = 0$ for $j = 0,\dots,N+1$, $j \neq i$. We call these functions the* hat functions *or* basis functions *for P_1 Lagrange interpolation.*

As we have said before, all functions in V_h are determined by their nodal values. In particular here, $w_h^i(a) = w_h^i(b) = 0$, since the endpoints correspond to $j = 0$ and $j = N+1$, and $1 \leq i \leq N$. The term P_1 Lagrange interpolation stems from the fact that affine functions are polynomials of degree at most 1, hence P_1, and that these functions are also used for Lagrange interpolation in V_h, as we will see shortly.

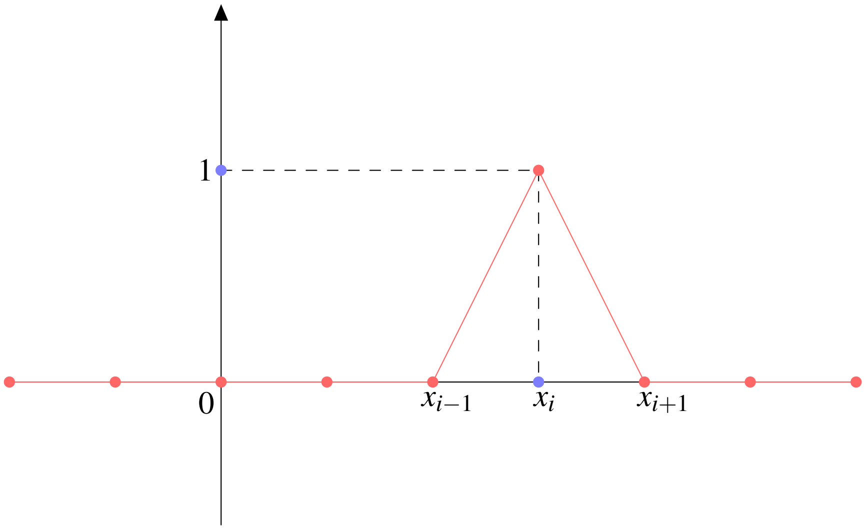

Figure 4. The hat function w_h^i with support $[x_{i-1}, x_{i+1}]$.

Proposition 4.2.3 *The family $(w_h^i)_{i=1,\dots,N}$ is a basis of V_h. Thus* $\dim V_h = N$. *Moreover, we have the* interpolation property

$$\forall v_h \in V_h, \quad v_h(x) = \sum_{i=1}^{N} v_h(x_i) w_h^i(x). \tag{4.5}$$

Proof. We use the Kronecker delta symbol: $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ otherwise. The hat functions thus satisfy $w_h^i(x_j) = \delta_{ij}$ for $i = 1,\dots,N$ and $j = 0,\dots,N+1$.

Let us first show that the family is linearly independent. Let λ_i be scalars such that

$$\sum_{i=1}^{N} \lambda_i w_h^i = 0.$$

Evaluating this zero function at point x_j yields

$$0 = \sum_{i=1}^{N} \lambda_i w_h^i(x_j) = \sum_{i=1}^{N} \lambda_i \delta_{ij} = \lambda_j$$

since in the last sum, the only nonzero term corresponds to $i = j$. Thus all coefficients vanish and the family is linearly independent.

Next we show that the family spans V_h. For all $v_h \in V_h$, we define

$$\widetilde{v}_h = \sum_{i=1}^{N} v_h(x_i) w_h^i \in V_h.$$

Now, of course $v_h - \widetilde{v}_h \in V_h$ and since $\widetilde{v}_h(x_j) = \sum_{i=1}^{N} v_h(x_i) w_h^i(x_j) = \sum_{i=1}^{N} v_h(x_i)\delta_{ij} = v_h(x_j)$ (same computation as above), then $(v_h - \widetilde{v}_h)(x_j) = 0$ for all $j = 0, \dots, N+1$. For each element $[x_j, x_{j+1}]$, we thus see that $v_h - \widetilde{v}_h$ is affine on the segment and vanishes at both endpoints, hence is identically zero on $[x_j, x_{j+1}]$. As this is true for all j, we have $v_h - \widetilde{v}_h = 0$ on $[a,b]$, that is to say $v_h = \widetilde{v}_h$, which shows both that the family is spanning and that we have formula (4.5).

The family $(w^i)_{i=1,\dots,N}$ is linearly independent and spanning, thus is a basis of V_h. It has N elements so that $\dim V_h = N$. □

Remark 4.2.3 The Lagrange interpolation property (4.5) is very important. It shows that with this specific choice of basis, the coordinates of a function v_h are precisely its nodal values $v_h(x_i)$. Hence solving the linear system $AX = B$ is going to directly provide the nodal values of the discrete solution u_h, without any post-processing. The linear forms $v_h \mapsto v_h(x_i)$, which belong to the dual V_h^* of V_h, are called the *degrees of freedom* in the FEM context. From the point of view of linear algebra, they are just the dual basis of the basis $(w^i)_{i=1,\dots,N}$. □

Corollary 4.2.1 *With the hat functions basis, the $N \times N$ matrix A is tridiagonal.*

Proof. Indeed, the support of w_h^i is $[x_{i-1}, x_{i+1}]$, therefore if $|i-j| \geq 2$, then $x_{i-1} \geq x_{j+1}$ or $x_{j-1} \geq x_{i+1}$ and the intersection of both supports is of zero measure, hence $A_{ij} = 0$. Thus, on any given line of the matrix A, we have at most three nonzero coefficients: $A_{i,i-1}$ corresponding to the subdiagonal, A_{ii} corresponding to the diagonal and $A_{i,i+1}$ corresponding to the superdiagonal. □

Of course, a tridiagonal matrix is the best kind of matrix that can be expected, apart from a diagonal matrix which cannot occur. This is because the problem is exceedingly simple. Actually, it is easy to compute all nonzero coefficients.

Proposition 4.2.4 *If $c = c_0$ and $f = f_0$ are constant, then we have*

$$A_{ii} = \frac{2}{h} + \frac{2h}{3}c_0, \quad A_{i,i-1} = A_{i,i+1} = -\frac{1}{h} + \frac{h}{6}c_0 \quad \textit{and} \quad B_i = hf_0.$$

Proof. We start by noticing that $w_h^i(x) = w_h^1(x - x_i)$ (extending w_h^1 by zero outside of $[a,b]$), thus $A_{ii} = A_{11}$ and $A_{i,i-1} = A_{i,i+1} = A_{12}$.

It is easy to see that $w_h^1(x) = \frac{x}{h}$ on $[x_0,x_1]$ and $w_h^1(x) = 2 - \frac{x}{h}$ on $[x_1,x_2]$, 0 elsewhere. Therefore, $(w_h^1)'(x) = \frac{1}{h}$ on $[x_0,x_1]$, $(w_h^1)'(x) = -\frac{1}{h}$ on $[x_1,x_2]$, 0 elsewhere. Thus

$$\begin{aligned}
A_{11} &= \int_{x_0}^{x_2} \big((w_h^1)'(x)^2 + c_0 w_h^1(x)^2\big)\,dx \\
&= \int_{x_0}^{x_1} \big((w_h^1)'(x)^2 + c_0 w_h^1(x)^2\big)\,dx + \int_{x_1}^{x_2} \big((w_h^1)'(x)^2 + c_0 w_h^1(x)^2\big)\,dx \\
&= \frac{1}{h^2} \times h + \frac{c_0}{h^2}\int_{x_0}^{x_1} x^2\,dx + \frac{1}{h^2} \times h + \frac{c_0}{h^2}\int_{x_1}^{x_2} (2h - x)^2\,dx \\
&= \frac{2}{h} + \frac{2h}{3}c_0.
\end{aligned}$$

The intersection of the supports of w_h^1 and w_h^2 is $[x_1,x_2]$, hence

$$\begin{aligned}
A_{12} &= \int_{x_1}^{x_2} \big((w_h^1)'(x)(w_h^2)'(x) + c_0 w_h^1(x) w_h^2(x)\big)\,dx \\
&= -\frac{1}{h^2} \times h + \frac{c_0}{h^2}\int_{x_1}^{x_2} (2h - x)(x - h)\,dx \\
&= -\frac{1}{h} + \frac{h}{6}c_0.
\end{aligned}$$

We leave the last value to the reader. □

Remark 4.2.4 When c or f are not constant, the corresponding terms may not necessarily be exactly computable and it may be necessary to resort to numerical integration. These terms however are corrections to the dominant terms $\frac{2}{h}$ and $-\frac{1}{h}$, so that it can be shown that choosing a sufficiently accurate numerical integration rule does not modify the final error estimate.

Numerical methods for linear systems are especially efficient in the case of a tridiagonal matrix. Among the classical methods used, let us mention *LU*, Cholesky (the matrix is symmetric), conjugate gradient. There are other more modern and sophisticated methods of course. □

4.3 A fourth order example

Let us now briefly consider the beam problem

$$\begin{cases} u^{(4)} + cu = f \text{ in } \Omega, \\ u(a) = u(b) = u'(a) = u'(b) = 0. \end{cases} \tag{4.6}$$

The variational formulation of this problem is set in $V = H_0^2(]a,b[)$ and since $H^2(]a,b[) \hookrightarrow C^1([a,b])$, the previous P_1 finite element method is not adapted (exercise, show that if v is piecewise affine, then $v'' = \sum_{i=1}^N (\lambda_i - \lambda_{i-1})\delta_{x_i} \notin L^2(a,b)$). We need higher degree polynomials to match not only the values, but the derivatives at mesh nodes. In the following, P_k denotes the space of polynomials of degree at most k. We thus define

$$V_h = \{v_h \in C^1([a,b]); v_{h|[x_i,x_{i+1}]} \in P_3 \text{ for } i = 0, \dots, N, \\ v_h(a) = v_h(b) = v_h'(a) = v_h'(b) = 0\}. \tag{4.7}$$

A natural question is why not simply use P_2 polynomials. The reason is that, in this case $V_h = \{0\}$ (exercise), which is not very good for approximation purposes. Degree $k = 3$ is the first degree for which the above definition gives rise to a nontrivial space, although we have no proof of this at present.

Proposition 4.3.1 *We have $V_h \subset H_0^2(]a,b[)$.*

Proof. Argue as in the proof of Proposition 4.2.1. □

We thus need C^1 functions, and in order to ensure the continuity and derivability at the endpoints of each element, we need to be able to specify both the value of the polynomial and of its derivative. The simplest way to achieve this is to use P_3 Hermite interpolation. Let us rapidly recall this interpolation.

Proposition 4.3.2 *For all quadruplets $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ of scalars, there exists a unique polynomial $P \in P_3$ such that*

$$P(0) = \alpha_0, \quad P(1) = \alpha_1, \quad P'(0) = \beta_0, \quad P'(1) = \beta_1.$$

The P_3 Hermite basis polynomials are given by

$$p_0(x) = (1-x)^2(1+2x), p_1(x) = x^2(3-2x), q_0(x) = x(1-x)^2, q_1(x) = x^2(x-1).$$

Proof. The proof of Proposition 4.3.2 follows from a simple dimension argument: we show that the linear mapping $P_3 \to \mathbb{R}^4$, $P \mapsto (P(0), P(1), P'(0), P'(1))$ is an isomorphism. Since P_3 is four-dimensional, it suffices to show that its kernel is trivial. But a polynomial such that $P(0) = P(1) = P'(0) = P'(1) = 0$ has a double root at $x = 0$ and another double root at $x = 1$, hence a number of roots counting multiplicities of at least four. We now that a nonzero polynomial of degree at most three has at most three roots. Hence $P = 0$. □

Of course, the P_3 Hermite basis polynomials form a basis of P_3, since they are the inverse image of the canonical basis of $\mathbb{R}^4$ by the previous isomorphism. They are uniquely determined by the following interpolation values: $(\alpha_0, \alpha_1, \beta_0, \beta_1) = (1,0,0,0)$ for p_0, $(\alpha_0, \alpha_1, \beta_0, \beta_1) = (0,1,0,0)$ for p_1, $(\alpha_0, \alpha_1, \beta_0, \beta_1) = (0,0,1,0)$ for q_0 and $(\alpha_0, \alpha_1, \beta_0, \beta_1) = (0,0,0,1)$ for q_1, so that any polynomial P of P_3 is uniquely written as

$$P = P(0)p_0 + P(1)p_1 + P'(0)q_0 + P'(1)q_1.$$

In FEM language, the linear forms $P \mapsto P(0)$, $P \mapsto P(1)$, $P \mapsto P'(0)$ and $P \mapsto P'(1)$ are the degrees of freedom of P_3 Hermite interpolation on the reference element $[0,1]$.

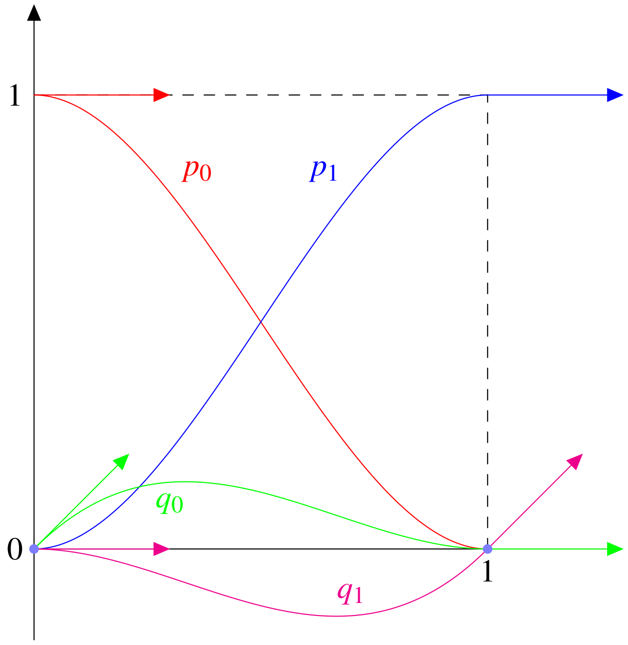

Figure 5. The four P_3 Hermite basis polynomials.

Once we have Hermite interpolation of the reference element $[0,1]$, we have Hermite interpolation on any element $[x_i, x_{i+1}]$ by a simple affine change of variables.

Lemma 4.3.1 *There exists a unique continuous linear mapping* $\Pi_h \colon H_0^2(]a,b[) \to V_h$*, again called the* V_h-interpolation operator *such that for all* v *in* $H_0^2(]a,b[)$*,* $v(x_i) = \Pi_h v(x_i)$ *and* $v'(x_i) = (\Pi_h v)'(x_i)$ *for* $i = 0, \dots, N+1$.

Proof. First of all, we note that $H_0^2(]a,b[) \hookrightarrow C^1([a,b])$, therefore the nodal values $v(x_i)$ and $v'(x_i)$ are unambiguously defined.

Now a P_3 polynomial on $[x_i, x_{i+1}]$ is uniquely determined by its values and the values of its derivative at x_i and x_{i+1}, by Hermite interpolation. Thus, the relations $v(x_i) = \Pi_h v(x_i)$ and $v'(x_i) = (\Pi_h v)'(x_i)$ for $i = 0, \dots, N+1$ define a unique piecewise P_3 function on the mesh, that is globally C^1 and vanishes at both ends together with its derivatices, thus belongs to V_h. Let $\Pi_h v$ be this function. Clearly, the mapping $v \mapsto \Pi_h v$ is linear from $H_0^1(]a,b[)$ into V_h. We leave the continuity to the reader. □

The above proof also shows that $V_h \neq \{0\}$. We can now show an error estimate, along the same lines as before.

Proposition 4.3.3 *Assume c and f are continuous on $[a,b]$. Then u is of class $C^4([a,b])$ and there exists a constant C independent of u such that*

$$\|u - u_h\|_V \leq Ch^2 \max_{[a,b]} |u^{(4)}|. \tag{4.8}$$

Proof. If f and c are continuous, since u is also continuous by Theorem 2.7.1, then $u^{(4)} = cu - f$ is continuous on $[a,b]$ and thus $u \in C^4([a,b])$. By Céa's lemma, we have

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \|u - \Pi_h u\|_V.$$

We use the H^2 semi-norm as a norm on V. We have

$$\|u - \Pi_h u\|_V^2 = \int_a^b ((u - \Pi_h u)'')^2 \, dx = \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} (u'' - (\Pi_h u)'')^2 \, dx.$$

Let us consider the function $w = u - \Pi_h u$ on $[x_i, x_{i+1}]$. By definition of V_h-interpolation, we have $w(x_i) = w(x_{i+1}) = 0$. Since w is C^1 on $[x_i, x_{i+1}]$, Rolle's theorem applies and there exists $c_1 \in]x_i, x_{i+1}[$ such that $w'(c_1) = 0$. Now, we also have $w'(x_i) = w'(x_{i+1}) = 0$ by V_h-interpolation, and w' is also of class C^1 so that Rolle applies again and there exists $c_2 < c_1 < c_3$ such that $w''(c_2) = w''(c_3) = 0$. We apply Rolle one last time since w'' is C^1 and obtain a point $c_4 \in [x_i, x_{i+1}]$ such that $w'''(c_4) = 0$. Consequently

$$w'''(x) = \int_{c_4}^{x} w^{(4)}(t) \, dt = \int_{c_4}^{x} u^{(4)} \, dt$$

for all $x \in [x_i, x_{i+1}]$, since $\Pi_h u$ is of degree at most three there, thus its fourth derivative vanishes. It follows from this equality that

$$|w'''(x)| \leq \int_{c_4}^{x} |u^{(4)}(t)| \, dt \leq \int_{x_i}^{x_{i+1}} |u^{(4)}(t)| \, dt \leq h \max_{[a,b]} |u^{(4)}|,$$

for all $x \in [x_i, x_{i+1}]$. We also have

$$w''(x) = \int_{c_2}^{x} w'''(t)\,dt,$$

so that substituting the previous estimate yields

$$|w''(x)| \leq h^2 \max_{[a,b]} |u^{(4)}|.$$

Squaring and integrating, we thus see that

$$\int_{x_i}^{x_{i+1}} (u'' - (\Pi_h u)'')^2\,dx = \int_{x_i}^{x_{i+1}} (w'')^2\,dx \leq h^5 \max_{[a,b]} |u^{(4)}|^2.$$

Now we sum from $i = 0$ to N

$$\|u - \Pi_h u\|_V^2 \leq h^5 \Big(\sum_{i=0}^{N} 1\Big) \max_{[a,b]} |u^{(4)}|^2 = h^4 (b-a) \max_{[a,b]} |u^{(4)}|^2,$$

which completes the proof. □

Remark 4.3.1 Under regularity hypotheses, we thus have convergence of the P_3 Hermite FEM based on the smallness of the interpolation error. It should be noted that this kind of proof relying on Rolle's theorem is not very natural in a Sobolev space context. There are better proofs using Hilbertian arguments. □

Let us say a few words about bases and matrices. It is apparent that the operator Π_h only uses the nodal values of the function and its derivatives. Hence, any set of interpolation data with N elements for the values and N elements for the derivative values gives rise to one and only one element of V_h. We thus define

Definition 4.3.1 *For $i = 1, \dots, N$, let $w_h^i \in V_h$ be defined by*

$$w_h^i(x_j) = \delta_{ij} \text{ and } (w_h^i)'(x_j) = 0,$$

and $z_h^i \in V_h$ be defined by

$$z_h^i(x_j) = 0 \text{ and } (z_h^i)'(x_j) = \delta_{ij},$$

for $j = 0, \dots, N+1$. We call these functions the basis functions *for P_3 Hermite interpolation on the mesh.*

The function w_h^i is thus equal to 1 at x_i and zero at all other nodes, with zero derivatives at all nodes, whereas the function z_h^i has derivative 1 at x_i and zero at all other nodes, with zero values at all nodes. Clearly they are constructed by pairing together the Hermite basis interpolation polynomials in each element $[x_{i-1}, x_i]$ and $[x_i, x_{i+1}]$, which are also called *shape functions* in the FEM context.

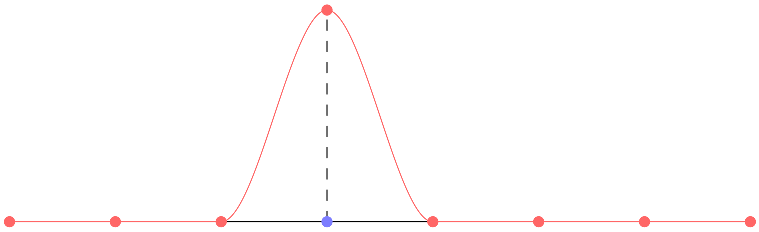

Figure 6a. A w_h^i basis function.

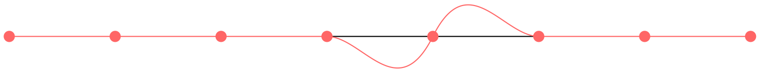

Figure 6b. A z_h^i basis function.

Proposition 4.3.4 *The family* $(w_h^i, z_h^i)_{i=1,\dots,N}$ *is a basis of* V_h. *Thus* $\dim V_h = 2N$. *Moreover, we have the* interpolation property

$$\forall v_h \in V_h, \quad v_h(x) = \sum_{i=1}^{N} v_h(x_i) w_h^i(x) + \sum_{i=1}^{N} v_h'(x_i) z_h^i(x). \tag{4.9}$$

Proof. Similar to the proof of Proposition 4.2.3 but using Hermite P_3 interpolation in each element. □

The supports of the basis functions are again $[x_{i-1}, x_{i+1}]$, thus we can expect lots of zero coefficients in the matrix. We do not write the detail here. Let us just mention that there is an issue of numbering. In the P_1 Lagrange case, there was a natural numbering of basis functions, which was that of the nodes. Here we have several choices, leading to different matrices. If we choose to number the basis elements as $(w_h^1, w_h^2, \dots, w_h^N, z_h^1, z_h^2, \dots, z_h^N)$, then the $2N \times 2N$ matrix A is comprised of four $N \times N$ blocks, and each one of the blocks is tridiagonal. If on the other hand, we interlace the basis functions like $(w_h^1, z_h^1, w_h^2, z_h^2, \dots, w_h^N, z_h^N)$, we obtain a sparse matrix each row of which has at most six nonzero coefficients grouped around the diagonal, resulting in seven nonzero diagonal rows: three above the diagonal and three under the diagonal.

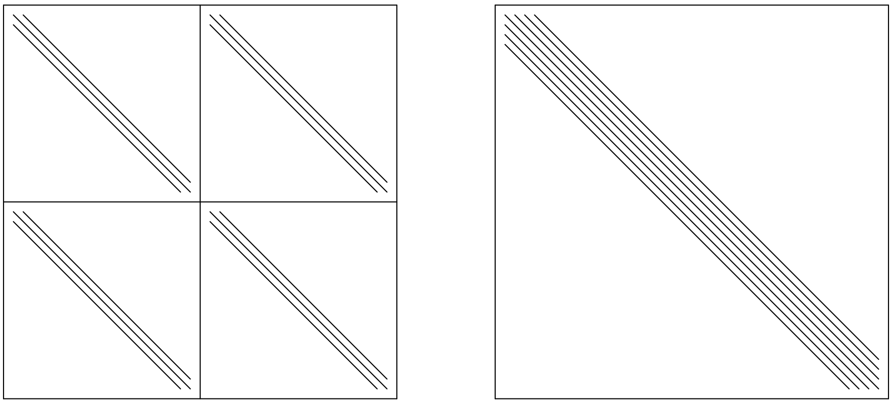

Figure 7. Matrix structure. Left: block tridiagonal, right: interlaced.

4.4 Neumann and Fourier conditions

Let us briefly indicate how to deal with Neumann and Fourier conditions for the model second order problem. There are several changes: the test-function space must not enforce boundary conditions, *i.e.*, $V = H^1(]a,b[)$, additional terms come up in the right hand-side for both problems, and there is an additional term in the bilinear form for the Fourier condition. Let us just consider the case $c > 0$. We thus let

$$V_h = \{v_h \in C^0([a,b]); v_{h|[x_i,x_{i+1}]} \in P_1\}. \tag{4.10}$$

Compared to the previous version of V_h, we have added two degrees of freedom $v_h \mapsto v(a)$ and $v_h \mapsto v(b)$, hence $\dim V_h = N+2$. We must accordingly complete the basis by adding two more basis functions w_h^0 and w_h^{N+1} defined by $w_h^0(x_j) = \delta_{0j}$ and $w_h^{N+1}(x_j) = \delta_{N+1,j}$ for all $j \in \{0,\dots,N+1\}$.

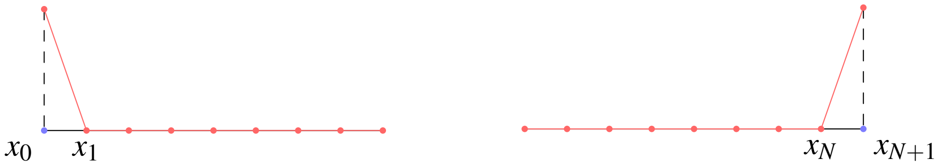

Figure 8. The two additional basis functions w_h^0 left and w_h^{N+1} right.

The variational formulation for the Fourier problem (replacing b by d in the Fourier condition to avoid a conflict in notation with the boundary b) is

$$\int_a^b (u'v' + cuv)\,dx + du(a)v(a) + du(b)v(b) = \int_a^b fv\,dx + gv(a) + gv(b).$$

We just set $d = 0$ for the Neumann problem. In matrix terms, the $(N+2) \times (N+2)$ matrix A is still symmetric tridiagonal, and we have

$$A_{00} = A_{N+1,N+1} = \frac{1}{h} + \frac{h}{3}c + d$$

and

$$A_{01} = A_{10} = A_{N,N+1} = A_{N+1,N} = -\frac{1}{h} + \frac{h}{6}c.$$

Of course

$$A_{0i} = A_{i0} = A_{j,N+1} = A_{N+1,j} = 0$$

for $i \geq 2$ and $j \leq N-1$. The other coefficients are unchanged. The right-hand side has two additional components $B_0 = B_{N+1} = \frac{hf}{2} + g$.

Similar changes must be made to treat a fourth order Neumann problem.