52
Piecewise linear approximation

> The results, however, of the labour and invention of this century are not to be found in a network of railways, in superb bridges, in enormous guns, or in instantaneous communication. We must compare the social state of the inhabitants of the country with what it was. The change is apparent enough. The population is double what it was a century back; the people are better fed and better housed, and comforts and even luxuries that were only within the reach of the wealthy can now be obtained by all classes alike.......But with these advantages there are some drawbacks. These have in many cases assumed national importance, and it has become the province of the engineer to provide a remedy. (Reynolds, 1868)

52.1 Introduction

Approximating a complicated function to arbitrary accuracy by "simpler" functions is a basic tool of applied mathematics. We have seen that piecewise polynomials are very useful for this purpose, and that is why approximation by piecewise polynomials plays a very important role in several areas of applied mathematics. For example, the *Finite Element Method* FEM is an extensively used tool for solving differential equations that is based on piecewise polynomial approximation, see the Chapters FEM for two-point boundary value problems and FEM for Poisson's equation.

In this chapter, we consider the problem of approximating a given real-valued function $f(x)$ on an interval $[a, b]$ by piecewise linear polynomials on a subdivision of $[a, b]$. We derive basic error estimates for interpolation with piecewise linear polynomials and we consider an application to least squares approximation.

52.2 Interpolation on $[0, 1]$

Let $f : [0, 1] \to \mathbb{R}$ be a given Lipschitz continuous function. Consider the function $\pi f : [0, 1] \to \mathbb{R}$ defined by

$$\pi f(x) = f(0)(1 - x) + f(1)x = f(0) + (f(1) - f(0))x.$$

Clearly, $\pi f(x)$ is a *linear* function in x,

$$\pi f(x) = c_0 + c_1 x,$$

where $c_0 = f(0)$, $c_1 = f(1) - f(0)$, and $\pi f(x)$ *interpolates* $f(x)$ at the end-points 0 and 1 of the interval $[0, 1]$, by which we mean that πf takes the same values as f at the end-points, i.e.

$$\pi f(0) = f(0), \quad \pi f(1) = f(1).$$

We refer to $\pi f(x)$ as a linear *interpolant* of $f(x)$ that interpolates $f(x)$ at the end-points of the interval $[0, 1]$.

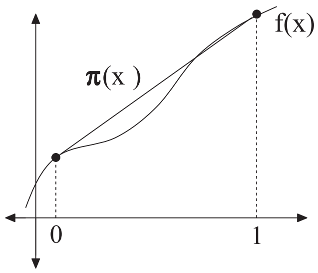

FIGURE 52.1. The linear interpolant πf of a function f.

We now study the *interpolation error* $f(x) - \pi f(x)$ for $x \in [0, 1]$. Before doing so we get some perspective on the space of linear functions on $[0, 1]$ to which the interpolant πf belongs.

52.3 The vector space of linear functions on $[0, 1]$

We let $\mathcal{P} = \mathcal{P}(0, 1)$ denote the set of first order (linear) polynomials

$$p(x) = c_0 + c_1 x,$$

defined for $x \in [0, 1]$, where the real numbers c_0 and c_1 are the coefficients of p. We recall that two polynomials $p(x)$ and $q(x)$ in $\mathcal{P}$ may be added to give a new polynomial $p + q$ in $\mathcal{P}$ defined by $(p + q)(x) = p(x) + q(x)$, and that a polynomial $p(x)$ in $\mathcal{P}$ may be multiplied by a scalar α to give a polynomial αp in $\mathcal{P}$ defined by $(\alpha p)(x) = \alpha p(x)$. Adding two polynomials is carried out by adding their coefficients, and multiplying a polynomial by a real number is carried out by multiplying the coefficients by the real number.

We conclude that $\mathcal{P}$ is a vector space where each vector is a particular first order polynomial $p(x) = c_0 + c_1 x$ determined by the two real numbers c_0 and c_1. As a basis for $\mathcal{P}$ we may choose $\{1, x\}$. To see this, we note that each $p \in \mathcal{P}$ can be uniquely expressed as a linear combination of 1 and x: $p(x) = c_0 + c_1 x$, and we may refer to the pair (c_0, c_1) as the coordinates of the polynomial $p(x) = c_0 + c_1 x$ with respect to the basis $\{1, x\}$. For example, the coordinates of the polynomial $p(x) = x$ with respect to the basis $\{1, x\}$, are $(0, 1)$, right? Since there are two basis functions, we say that the dimension of the vector space $\mathcal{P}$ is equal to two.

We now consider an alternative basis $\{\lambda_0, \lambda_1\}$ for $\mathcal{P}$ consisting of the two functions λ_0 and λ_1 defined

$$\lambda_0(x) = 1 - x, \quad \lambda_1(x) = x.$$

Each of these functions takes the value 0 at one end-point and the value 1 at the other end-point, namely

$$\lambda_0(0) = 1,\ \lambda_0(1) = 0, \qquad \text{and } \lambda_1(0) = 0, \lambda_1(1) = 1.$$

See Fig. 52.2.

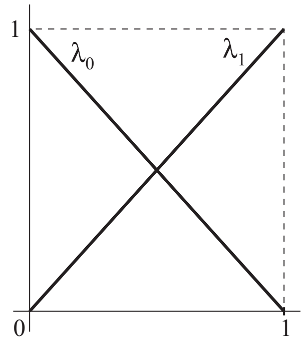

FIGURE 52.2. The basis functions λ_0 and λ_1.

Any polynomial $p(x) = c_0 + c_1 x$ in $\mathcal{P}$ can be expressed as a linear combination of the functions $\lambda_0(x)$ and $\lambda_1(x)$, i.e.

$$\begin{aligned} p(x) &= c_0 + c_1 x = c_0(1 - x) + (c_1 + c_0)x = c_0\lambda_0(x) + (c_1 + c_0)\lambda_1(x) \\ &= p(0)\lambda_0(x) + p(1)\lambda_1(x), \end{aligned}$$

A very nice feature of these functions is that the coefficients $p(0)$ and $p(1)$ are the values of $p(x)$ at $x = 0$ and $x = 1$. Moreover, λ_0 and λ_1 are linearly independent, since if

$$a_0\lambda_0(x) + a_1\lambda_1(x) = 0 \quad \text{for } x \in [0, 1],$$

then setting $x = 0$ and $x = 1$ shows that $a_1 = a_0 = 0$. We conclude that $\{\lambda_0, \lambda_1\}$ is a basis for $\mathcal{P}$.

In particular, we can express the interpolant $\pi f \in \mathcal{P}$ in the basis $\{\lambda_0, \lambda_1\}$ as follows:

$$\pi f(x) = f(0)\lambda_0(x) + f(1)\lambda_1(x), \tag{52.1}$$

where the end-point values $f(0)$ and $f(1)$ appear as coefficients.

52.3.1 The interpolation error

We want to estimate the interpolation error $f(x) - \pi f(x)$ for $x \in [0, 1]$. We prove that

$$|f(x) - \pi f(x)| \leq \frac{1}{2}x(1-x) \max_{y\in[0,1]} |f''(y)|, \quad x \in [0, 1]. \tag{52.2}$$

Since (convince yourself!)

$$0 \leq x(1-x) \leq \frac{1}{4} \quad \text{for } x \in [0, 1],$$

we can state the interpolation error estimate in the form

$$\max_{x\in[0,1]} |f(x) - \pi f(x)| \leq \frac{1}{8} \max_{y\in[0,1]} |f''(y)|. \tag{52.3}$$

This estimate states that the maximal value of the interpolation error $|f(x) - \pi f(x)|$ over $[0, 1]$ is bounded by a constant times the maximum value of the second derivative $|f''(y)|$ over $[0, 1]$, i.e. to the degree of concavity or convexity of f, or the amount that f curves away from being linear, see Fig. 52.3.

To prove (52.2), we fix x in $(0, 1)$ and use Taylor's theorem to express the values $f(0)$ and $f(1)$ in terms of $f(x)$, $f'(x)$, $f''(y_0)$ and $f''(y_1)$ where $y_0 \in (0, x)$ and $y_1 \in (x, 1)$. This gives

$$\begin{aligned} f(0) &= f(x) + f'(x)(-x) + \frac{1}{2}f''(y_0)(-x)^2, \\ f(1) &= f(x) + f'(x)(1-x) + \frac{1}{2}f''(y_1)(1-x)^2. \end{aligned} \tag{52.4}$$

Substituting the Taylor expansions (52.4) into (52.1) and using the identities

$$\begin{aligned} \lambda_0(x) + \lambda_1(x) &= (1-x) + x \equiv 1, \\ (-x)\lambda_0(x) + (1-x)\lambda_1(x) &= (-x)(1-x) + (1-x)x \equiv 0, \end{aligned} \tag{52.5}$$

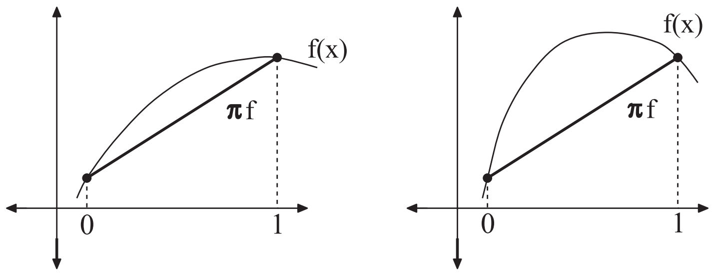

FIGURE 52.3. The error of a linear interpolant depends on the size of $|f''|$, which measures the degree that f curves away from being linear. Notice that the error of the linear interpolant of the function on the right is much larger than of the linear interpolant of the function on the left and the function on the right has a larger second derivative in magnitude.

we obtain the *error representation*

$$f(x) - \pi f(x) = -\frac{1}{2}\big(f''(y_0)(-x)^2(1-x) + f''(y_1)(1-x)^2 x\big).$$

Using the identity $(-x)^2(1-x)+(1-x)^2x = x(1-x)(x+1-x) = x(1-x)$ gives (52.2),

$$|f(x) - \pi f(x)| \leq \frac{1}{2}x(1-x)\max_{y\in[0,1]}|f''(y)| \leq \frac{1}{8}\max_{y\in[0,1]}|f''(y)|. \qquad (52.6)$$

Next, we prove the following estimate for the error in the first derivative,

$$|f'(x) - (\pi f)'(x)| \leq \frac{x^2+(1-x)^2}{2}\max_{y\in[0,1]}|f''(y)|, \quad x \in [0,1]. \qquad (52.7)$$

Since $0 \leq x^2 + (1-x)^2 \leq 1$ for $x \in [0, 1]$,

$$\max_{x\in[0,1]}|f'(x) - (\pi f)'(x)| \leq \frac{1}{2}\max_{y\in[0,1]}|f''(y)|.$$

We illustrate in Fig. 52.4.

To prove (52.7), we differentiate (52.1) with respect to x (note that the x-dependence is carried by $\lambda_0(x)$ and $\lambda_1(x)$) and use (52.4) together with the obvious identities

$$\begin{aligned} \lambda_0'(x) + \lambda_1'(x) &= -1 + 1 \equiv 0, \\ (-x)\lambda_0'(x) + (1-x)\lambda_1'(x) &= (-x)(-1) + (1-x) \equiv 1. \end{aligned}$$

This gives the error representation:

$$f'(x) - (\pi f)'(x) = -\frac{1}{2}\big(f''(y_0)(-x)^2(-1) + f''(y_1)(1-x)^2\big),$$

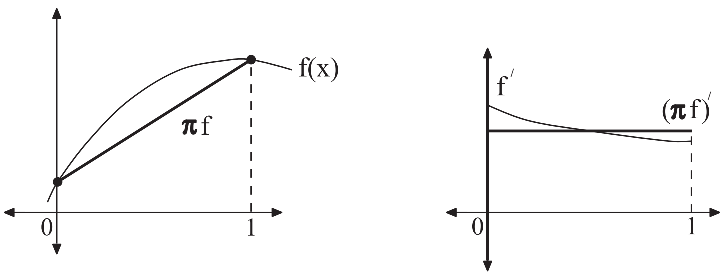

FIGURE 52.4. The derivative of a linear interpolant of f approximates the derivative of f. We show f and the linear interpolant πf on the left and their derivatives on the right.

where again $y_0 \in (0, x)$ and $y_1 \in (x, 1)$. This proves the desired result.

Finally, we prove an estimate for $|f(x) - \pi f(x)|$ using only the first derivative f'. This is useful when the second derivative f'' does not exist. The Mean Value theorem implies

$$f(0) = f(x) + f'(y_0)(-x), \quad f(1) = f(x) + f'(y_1)(1-x), \tag{52.8}$$

where $y_0 \in [0, x]$ and $y_1 \in [x, 1]$. Substituting into (52.1), we get

$$|f(x) - \pi f(x)| = |f'(y_0)x(1-x) - f'(y_1)(1-x)x| \leq 2x(1-x) \max_{y\in[0,1]} |f'(y)|.$$

Since $2x(1-x) \leq \frac{1}{2}$ for $0 \leq x \leq 1$, we thus find that

$$\max_{x\in[0,1]} |f(x) - \pi f(x)| \leq \frac{1}{2} \max_{y\in[0,1]} |f'(y)|.$$

We summarize in the following theorem.

Theorem 52.1 *The linear polynomial $\pi f \in \mathcal{P}(0,1)$, which interpolates the given function $f(x)$ at $x = 0$ and $x = 1$, satisfies the following error bounds:*

$$\begin{aligned} \max_{x\in[0,1]} |f(x) - \pi f(x)| &\leq \frac{1}{8} \max_{y\in[0,1]} |f''(y)|, \\ \max_{x\in[0,1]} |f(x) - \pi f(x)| &\leq \frac{1}{2} \max_{y\in[0,1]} |f'(y)|, \\ \max_{x\in[0,1]} |f'(x) - (\pi f)'(x)| &\leq \frac{1}{2} \max_{y\in[0,1]} |f''(y)|. \end{aligned} \tag{52.9}$$

The corresponding estimates for an arbitrary interval $I = [a, b]$ of length $h = b - a$ takes the following form, where of course $\mathcal{P}(a, b)$ denotes the set of linear functions on $[a, b]$. Observe how the length $h = b - a$ of the interval enters, with the factor h^2 in the estimate for $f(x) - \pi f(x)$ with f'', and h in the estimate for $f'(x) - (\pi f)'(x)$.

Theorem 52.2 *The linear polynomial $\pi f \in \mathcal{P}(a,b)$, which interpolates the given function $f(x)$ at $x = a$ and $x = b$, satisfies the following error bounds:*

$$\begin{aligned}\max_{x\in[a,b]} |f(x) - \pi f(x)| &\le \frac{1}{8} \max_{y\in[a,b]} |h^2 f''(y)|,\\ \max_{x\in[a,b]} |f(x) - \pi f(x)| &\le \frac{1}{2} \max_{y\in[a,b]} |h f'(y)|,\\ \max_{x\in[a,b]} |f'(x) - (\pi f)'(x)| &\le \frac{1}{2} \max_{y\in[a,b]} |h f''(y)|,\end{aligned} \tag{52.10}$$

where $h = b - a$.

If we define the *maximum norm* over $I = [a,b]$ by

$$\|v\|_{L_\infty(I)} = \max_{x\in[a,b]} |v(x)|,$$

then we can state (52.9) as follows

$$\begin{aligned}\|f - \pi f\|_{L_\infty(I)} &\le \frac{1}{8} \|h^2 f''\|_{L_\infty(I)},\\ \|f - \pi f\|_{L_\infty(I)} &\le \frac{1}{2} \|h f'\|_{L_\infty(I)},\\ \|f' - (\pi f)'\|_{L_\infty(I)} &\le \frac{1}{2} \|h f''\|_{L_\infty(I)}.\end{aligned} \tag{52.11}$$

Below we shall use an analog of this estimate with the $L_\infty(I)$-norm replaced by the $L_2(I)$-norm.

52.4 The vector space of piecewise linear continuous functions on an interval

For a given interval $I = [a,b]$, we let $a = x_0 < x_1 < x_2 < \cdots < x_N = b$ be a *partition* of I into N sub-intervals $I_i = (x_{i-1}, x_i)$ of length $h_i = x_i - x_{i-1}$, $i = 1, ..., N$. We denote by $h(x)$ the *mesh function* defined by $h(x) = h_i$ for $x \in I_i$ and we use $\mathcal{T}_h = \{I_i\}_{i=1}^N$ to denote the set of intervals or *mesh* or partition.

We introduce the vector space V_h of continuous piecewise linear functions on the mesh $\mathcal{T}_h$. A function $v \in V_h$ is linear on each subinterval I_i and is continuous on $[a,b]$. Adding two functions in V_h or multiplying a function in V_h by a real number gives a new function in V_h, and thus V_h is indeed a vector space. We show an example of such a function in Fig. 53.2.

We now present a particularly important basis for V_h that consists of the *hat functions* or *nodal basis functions* $\{\varphi_i\}_{i=0}^N$ illustrated in Fig. 52.5. The hat-function $\varphi_i(x)$ is a function in V_h satisfying

$$\varphi_i(x_j) = 1 \quad \text{if } j = i, \quad \varphi_i(x_j) = 0 \quad \text{if } j \ne i.$$

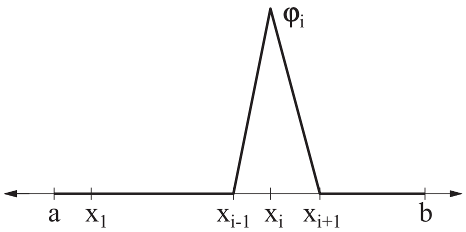

FIGURE 52.5. The hat function φ_i associated to node x_i.

and is given by the formula:

$$\varphi_i(x) = \begin{cases} 0, & x \notin [x_{i-1}, x_{i+1}], \\ \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i], \\ \dfrac{x - x_{i+1}}{x_i - x_{i+1}}, & x \in [x_i, x_{i+1}]. \end{cases}$$

The basis functions φ_0 and φ_N associated to the boundary nodes x_0 and x_N look like "half hats". Observe that each hat function $\varphi_i(x)$ is defined on the whole interval $[a, b]$ and takes the value zero outside the interval $[x_{i-1}, x_{i+1}]$ (or $[a, x_1]$ if $i = 0$ and $[x_{N-1}, b]$ if $i = N$).

The set of hat-functions $\{\varphi_i\}_{i=0}^N$ is a basis for V_h because each $v \in V_h$ has the unique representation

$$v(x) = \sum_{i=0}^{N} v(x_i)\varphi_i(x),$$

where the nodal values $v(x_i)$ appear as coefficients. To see this, it is sufficient to realize that the functions on the left and right hand side are both continuous and piecewise linear and take the same values at the nodes, and thus coincide. Since the number of basis functions φ_i is equal to $N + 1$, the dimension of V_h is equal to $N + 1$.

The continuous piecewise linear interpolant $\pi_h f \in V_h$ of a given Lipschitz continuous function $f(x)$ on $[0, 1]$ is defined by

$$\pi_h f(x_i) = f(x_i) \quad \text{for } i = 0, 1, ..., N,$$

that is, $\pi_h f(x)$ interpolates $f(x)$ at the nodes x_i, see Fig. 52.6. We can express $\pi_h f$ in terms of the basis of hat functions $\{\varphi_i\}_{i=0}^N$ as follows:

$$\pi_h f = \sum_{i=0}^{N} f(x_i)\varphi_i \quad \text{or} \quad \pi_h f(x) = \sum_{i=0}^{N} f(x_i)\varphi_i(x) \quad \text{for } x \in [0, 1], \tag{52.12}$$

with the x-dependence indicated.

Since $\pi_h f(x)$ is linear on each subinterval I_i and interpolates $f(x)$ at the end-points of I_i, we can express $\pi f(x)$ analytically on I_i as follows:

$$\pi_h f(x) = f(x_{i-1})\frac{x - x_i}{x_{i-1} - x_i} + f(x_i)\frac{x - x_{i-1}}{x_i - x_{i-1}} \quad \text{for } x_{i-1} \le x \le x_i,$$

for $i = 1, ..., N$.

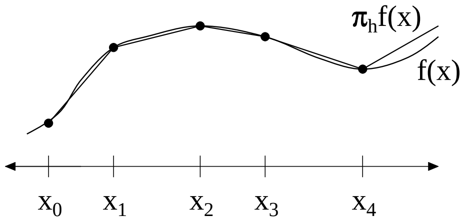

FIGURE 52.6. An example of a continuous piecewise linear interpolant.

Using Theorem 52.2, we obtain the following error estimate for piecewise linear interpolation:

Theorem 52.3 *The piecewise linear interpolant $\pi_h f(x)$ of a twice differentiable function $f(x)$ on a partition of $[a, b]$ with mesh function $h(x)$ satisfies*

$$\begin{aligned} \|f - \pi_h f\|_{L_\infty(a,b)} &\le \frac{1}{8}\|h^2 f''\|_{L_\infty(a,b)}, \\ \|f' - (\pi_h f)'\|_{L_\infty(a,b)} &\le \frac{1}{2}\|h f''\|_{L_\infty(a,b)}. \end{aligned} \tag{52.13}$$

If $f(x)$ is only once differentiable, then

$$\|f - \pi_h f\|_{L_\infty(a,b)} \le \frac{1}{2}\|h f'\|_{L_\infty(a,b)}. \tag{52.14}$$

Note that since the mesh function $h(x)$ may have jumps at the nodes, we interpret $\|h^2 f''\|_{L_\infty(a,b)}$ as

$$\max_{i=1,\ldots,N} \max_{y\in[x_{i-1},x_i]} |h^2(y)\, f''(y)|,$$

where $h(y) = x_i - x_{i-1}$ for $y \in [x_{i-1}, x_i]$.

52.5 The L_2 projection into the space V_h of piecewise linear functions on I

Let $f(x)$ be a given function on an interval $I = [a, b]$ and V_h denote the space of continuous piecewise linear functions V_h on a partition $a = x_0 < ... < x_N = b$ of I with mesh function $h(x)$.

The *orthogonal projection* $P_h f$ of the function f into V_h is the function $P_h f \in V_h$ such that

$$\int_I (f - P_h f) v \, dx = 0 \quad \text{for } v \in V_h. \tag{52.15}$$

Recalling the definition of the $L_2(I)$-scalar product

$$(v, w)_{L_2(I)} = \int_I v(x) w(x) \, dx,$$

with the corresponding $L_2(I)$-norm

$$\|v\|_{L_2(I)} = (\int_I v^2(x) \, dx)^{1/2},$$

we can write (52.15) in the form

$$(f - P_h f, v)_{L_2(I)} = 0 \quad \text{for } v \in V_h.$$

This says that $f - P_h f$ is orthogonal to V_h with respect to the $L_2(I)$ scalar product. We also call $P_h f$ the $L_2(I)$-*projection* of f onto V_h.

We first show that $P_h f$ is uniquely defined and then prove that $P_h f$ is the best V_h-approximation of f in the $L_2(I)$-norm.

To prove uniqueness and existence, we express $P_h f$ in the nodal basis $\{\varphi_i\}_{i=0}^N$:

$$P_h f(x) = \sum_{j=0}^N c_j \varphi_j(x),$$

where the $c_j = (P_h f)(x_j)$ are the nodal values of $P_h f$ that have to be determined. We insert this representation into (52.15) and choose $v = \varphi_i$ with $i = 0, ..., N$, to get

$$\begin{aligned}\int_I \sum_{j=0}^N c_j \varphi_j(x) \varphi_i(x) \, dx &= \sum_{j=0}^N c_j \int_I \varphi_j(x) \varphi_i(x) \, dx \\ &= \int_I f \varphi_i \, dx \equiv b_i \quad i = 0, 1, ..., N, \end{aligned} \tag{52.16}$$

where we changed the order of integration and summation. This gives the following system of equations

$$\sum_{j=0}^N m_{ij} c_j = \int_I f \varphi_i \, dx \equiv b_i \quad i = 0, 1, ..., N, \tag{52.17}$$

where

$$m_{ij} = \int_I \varphi_j(x) \varphi_i(x) \, dx, \quad i, j = 0, ..., N.$$

We can write (52.17) in matrix form as

$$Mc = b$$

where $c = (c_0, ..., c_N)$ is a $N+1$-vector of the unknown coefficients c_j, and $b = (b_0, ..., b_N)$ is computable from $f(x)$, and $M = (m_{ij})$ is a $(N+1)\times(N+1)$-matrix that depends on the basis functions φ_i, but not on the function $f(x)$. We refer to the matrix M as the *mass matrix*.

We can now easily prove the uniqueness of $P_h f$. Since the difference $P_h f - \bar{P}_h f$ of two functions $P_h f \in V_h$ and $\bar{P}_h f \in V_h$ satisfying the relation (52.15), also satisfy

$$\int_I (P_h f - \bar{P}_h f) v \, dx = 0 \quad \text{for } v \in V_h,$$

by choosing $v = P_h f - \bar{P}_h f$, we get

$$\int_I (P_h f - \bar{P}_h f)^2 \, dx = 0,$$

and thus $P_h f(x) = \bar{P}_h f(x)$ for $x \in I$. Solutions of the system $Mc = b$ are therefore unique, and since M is a square matrix, existence follows from the Fundamental Theorem of Linear Algebra. We sum up:

Theorem 52.4 *The $L_2(I)$-projection $P_h f$ of a given function f onto the set of piecewise linear functions V_h on I is uniquely defined by* (52.15) *or the equivalent system of equations $Mc = b$, where $c_j = P_h f(x_j)$ are the nodal values of $P_h f$, M is the mass matrix with coefficients $m_{ij} = (\varphi_j, \varphi_i)_{L_2(I)} = (\varphi_i, \varphi_j)_{L_2(I)}$ and the coefficients of the right hand side b are given by $b_i = (f, \varphi_i)$.*

EXAMPLE 52.1. We compute the mass matrix M in the case of a uniform subdivision with $h(x) = h = (b-a)/N$ for $x \in I$. We get by a direct computation

$$m_{ii} = \int_{x_{i-1}}^{x_{i+1}} \varphi_i^2(x) \, dx = \frac{2h}{3} \quad i = 1, ...N-1, \qquad m_{00} = m_{NN} = \frac{h}{3},$$

$$m_{i,i+1} = \int_{x_{i-1}}^{x_{i+1}} \varphi_i(x)\varphi_{i+1}(x) \, dx = \frac{h}{6} \quad i = 1, ...N-1.$$

The corresponding "lumped" mass matrix $\hat{M} = (\hat{m}_{ij})$, which is a diagonal matrix with the diagonal element in each row being the sum of the elements in the corresponding row of M, takes the form

$$\hat{m}_{ii} = h \quad i = 1, ..., N-1, \quad \hat{m}_{00} = \hat{m}_{NN} = h/2.$$

We see that $\hat{M}$ may be viewed as a h-scaled variant of the identity matrix and M can be viewed as an h-scaled approximation of the identity matrix.

We now prove that the $L_2(I)$-projection $P_h f$ of a function f satisfies

$$\|f - P_h f\|_{L_2(I)} \le \|f - v\|_{L_2(I)}, \quad \text{for all } v \in V_h. \tag{52.18}$$

This implies that $P_h f$ is the element in V_h with smallest deviation from f in the $L_2(I)$-norm. Applying Cauchy's inequality to (52.15) with $v \in V_h$ gives

$$\begin{aligned}
&\int_I (f - P_h f)^2\,dx \\
&= \int_I (f - P_h f)(f - P_h f)\,dx + \int_I (f - P_h f)(P_h f - v)\,dx \\
&= \int_I (f - P_h f)(f - v)\,dx \le \Big(\int_I (f - P_h f)^2\,dx\Big)^{1/2}\Big(\int_I (f - v)^2\,dx\Big)^{1/2},
\end{aligned}$$

which proves the desired result. We summarize:

Theorem 52.5 *The $L_2(I)$-projection P_h into V_h defined by* (52.15)*, is the unique element in V_h which minimizes $\|f - v\|_{L_2(I)}$ with v varying over V_h.*

In particular, choosing $v = \pi_h f$ in (52.18), we obtain

$$\|f - P_h f\|_{L_2(I)} \le \|f - \pi_h f\|_{L_2(I)},$$

where $\pi_h f$ is the nodal interpolant of f introduced above. One can prove the following analog of (52.13)

$$\|f - \pi_h f\|_{L_2(I)} \le \frac{1}{\pi^2}\|h^2 f''\|_{L_2(I)},$$

where the interpolation constant happens to be π^{-2}. We thus conclude the following basic result:

Theorem 52.6 *The $L_2(I)$-projection P_h into the space of piecewise linear functions V_h on I with mesh function $h(x)$, satisfies the following error estimate:*

$$\|f - P_h f\|_{L_2(I)} \le \frac{1}{\pi^2}\|h^2 f''\|_{L_2(I)}. \tag{52.19}$$

Chapter 52 Problems

52.1. Give a different proof of the first estimate of Theorem Theorem 52.1 by considering for a given $x \in (0,1)$, the function

$$g(y) = f(y) - \pi f(y) - \gamma(x) y(1-y), \quad y \in [0,1],$$

where $\gamma(x)$ is chosen so that $g(x) = 0$. Hint: the function $g(y)$ vanishes at 0, x and 1. Show by repeated use of the Mean Value theorem that g'' vanishes at some point ξ, from which it follows that $\gamma(x) = -f''(\xi)/2$.

52.2. Prove Theorem 52.2 from Theorem 52.1 by using the change of variables $x = a + (b-a)z$ transforming the interval $[0,1]$ onto $[a,b]$, setting $F(z) = f(a + (b-a)z)$ and using that by the Chain Rule, $F' = \frac{dF}{dz} = (b-a)f' = (b-a)\frac{df}{dx}$.

52.3. Develop approximation/interpolation with piecewise constant (discontinuous) functions on a partition of an interval. Consider interpolation at left-hand endpoint, right-hand endpoint, midpoint and mean value for each subinterval. Prove error estimates of the form $\|u - \pi_h u\|_{L_\infty(I)} \leq C\|hu'\|_{L_\infty(I)}$, with $C = 1$ or $C = \frac{1}{2}$.

53
FEM for two-point boundary value problems

> The beginners mind is empty, free of the habits of the expert, ready to accept, or doubt, and open to all the possibilities. It is a kind of mind which can see things as they are. (Shunryu Suzuki)

53.1 Introduction

We begin by deriving a model that is based on a *conservation principle* which states:

> The rate at which a specified quantity changes in a region is equal to the rate that the quantity leaves and enters the region plus the rate at which the quantity is created and destroyed inside the region.

Such a conservation principle holds for a wide variety of quantities, including animals, automobiles, bacteria, chemicals, fluids, heat and energy, etc. So the model we derive has a wide application.

In this chapter, we assume that the quantity to be modeled exists in a very small diameter "tube" with constant cross section and that the quantity varies in the direction along the tube but not at all within a fixed cross section, see Fig. 53.1. We use x to denote the position along the length of the tube and let t denote time. We assume that the quantity in the tube is sufficiently abundant that it makes sense to talk about a *density* $u(x,t)$, measured in amount of the quantity per unit volume, that varies continuously with the position x and time t. This is certainly valid for quantities such as heat and energy, and may be more or less valid for quantities such as bacteria and chemicals provided there is a sufficient number of creatures or molecules respectively.

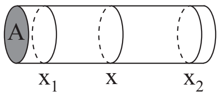

FIGURE 53.1. Variation in a very narrow "tube".

We next express the conservation principle mathematically. We consider a small region of the tube of width dx and cross-sectional area A. The amount of quantity in this region is $u(x,t)A\,dx$. We let $q(x,t)$ denote the *flux* at position x and time t, or the amount of the quantity crossing the section at x at time t measured in amount per unit area per unit time. We choose the orientation so that q is positive when the flow is to the right. The amount of quantity crossing the section at position x at time t is therefore $Aq(x,t)$. Lastly, we let $f(x,t)$ denote the rate that the quantity is created or destroyed within the section at x at time t measured in amount per unit volume per unit time. So, $f(x,t)Adx$ is the amount of the quantity created or destroyed in the small region of width dx per unit time.

The conservation principle for a fixed length of pipe between $x = x_1$ and $x = x_2$ implies that the rate of change of the quantity in this section must equal the rate at which it flows in at $x = x_1$ minus the rate at which it flows out at $x = x_2$ plus the rate at which it is created in $x_1 \leq x \leq x_2$. In mathematical terms,

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)A\,dx = Aq(x_1,t) - Aq(x_2,t) + \int_{x_1}^{x_2} f(x,t)A\,dx.$$

or

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)\,dx = q(x_1,t) - q(x_2,t) + \int_{x_1}^{x_2} f(x,t)\,dx. \tag{53.1}$$

Equation (53.1) is called the *integral formulation* of the conservation principle.

We can reformulate (53.1) as a partial differential equation provided $u(x,t)$ and $q(x,t)$ are sufficiently smooth. For we can write,

$$\begin{aligned}\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)\,dx &= \int_{x_1}^{x_2} \frac{\partial}{\partial t}u(x,t)\,dx,\\ q(x_1,t) - q(x_2,t) &= \int_{x_1}^{x_2} \frac{\partial}{\partial x}q(x,t)\,dx,\end{aligned}$$

and therefore collecting terms,

$$\int_{x_1}^{x_2} \big(\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}q(x,t) - f(x,t)\big)\,dx = 0.$$

Since x_1 and x_2 are arbitrary, the integrand must be zero at each point, or

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}q(x,t) = f(x,t). \tag{53.2}$$

Equation (53.2) is the *pointwise* or *differential formulation* of the conservation principle.

So far we have one equation for two unknowns. To complete the model, we use a *constitutive relation* that describes the relation between the flux and the density. This relation is specific to the physical properties of the quantity being modeled, yet it is often unclear exactly how to model these properties. A constitutive relation used in practice is often only an approximation to the true unknown relation.

Many quantities have the property that the quantity flows from regions of high concentration to regions of low concentration, and the rate of flow increases as the differences in concentration increases. As a first approximation, we assume a simple linear relation

$$q(x,t) = -a(x,t)\frac{\partial}{\partial x}u(x,t), \tag{53.3}$$

where $a(x,t) > 0$ is the *diffusion coefficient.* In case u represents heat, (53.3) is known as *Newton's Heat Law.* In general, equation (53.3) is known as *Fick's Law.* Note that the choice of sign of a guarantees for example that flow is to the right if $u_x < 0$, i.e. if u decreases across the section at x. Substituting (53.3) into (53.2), we obtain the general time-dependent reaction-diffusion equation,

$$\frac{\partial}{\partial t}u(x,t) - \frac{\partial}{\partial x}\big(a(x,t)\frac{\partial}{\partial x}u(x,t)\big) = f(x,t).$$

To simplify the notation, we use $\dot{u}$ to denote $\partial u/\partial t$ and u' to denote $\partial u/\partial x$. This yields

$$\dot{u}(x,t) - (a(x,t)u'(x,t))' = f(x,t). \tag{53.4}$$

Convection or transport is another important process to take into account in this model.

EXAMPLE 53.1. When modeling populations of animals, diffusion reflects the natural tendency of most creatures to spread out over a region due to randomly occurring interactions between pairs of creatures, while convection models phenomena such as migration.

Convection is modeled by assuming a constitutive relation in which the flux is proportional to the density, i.e.

$$\varphi(x,t) = b(x,t)u(x,t),$$

which results in a convection term in the differential equation of the form $(bu)'$. The convection coefficient $b(x,t)$ determines the rate and direction of transport of the quantity being modeled.

In general, many quantities are modeled by a constitutive relation of the form

$$\varphi(x,t) = -a(x,t)u'(x,t) + b(x,t)u(x,t)$$

which combines diffusion and convection. Arguing as above, we obtain the general reaction-diffusion-convection equation

$$\dot{u}(x,t) - (a(x,t)u'(x,t))' + (b(x,t)u(x,t))' = f(x,t). \tag{53.5}$$

53.2 Initial boundary-value problems

We have to add suitable data to (53.4) or (53.5) in order to specify a unique solution. We model the amount of substance in a fixed length of tube located between $x = 0$ and $x = 1$, as in Fig. 53.1, and specify some information about u called *boundary conditions* at $x = 0$ and $x = 1$. We also need to give some initial data at some initial time, which we take to be $t = 0$. The *evolutionary* or *time-dependent initial two point boundary value problem* reads: find $u(x,t)$ such that

$$\begin{cases} \dot{u} - (au')' + (bu)' = f & \text{in } (0,1)\times(0,T), \\ u(0,t) = u(1,t) = 0 & \text{for } t \in (0,T) \\ u(x,0) = u_0(x) & \text{for } x \in (0,1), \end{cases} \tag{53.6}$$

where a, b, c are given coefficients and f and g are given data. The boundary values $u(0,t) = u(1,t) = 0$ are known as *homogeneous Dirichlet boundary conditions.*

EXAMPLE 53.2. In the case that we use (53.4) to model the heat u in a long thin wire, the coefficient a represents the heat conductivity of the metal in the wire, f is a given heat source, and the homogeneous Dirichlet boundary conditions at the end-points means that the temperature of the wire is held fixed at 0 there. Such conditions are realistic for example if the wire is attached to very large masses at the ends.

Other boundary conditions found in practice include: *nonhomogeneous Dirichlet boundary conditions* $u(0) = u_0$, $u(1) = u_1$ with constants u_0, u_1; one homogeneous Dirichlet $u(0) = 0$ and one *nonhomogeneous Neumann boundary condition* $a(1)u'(1) = g_1$ with constant g_1; and more general *Robin boundary conditions*

$$-a(0)u'(0) = \gamma(0)(u_0 - u(0)), \quad a(1)u'(1) = \gamma(1)(u_1 - u(1))$$

with constants $\gamma(0)$, u_0, $\gamma(1)$, u_1.

53.3 Stationary boundary value problems

In many situations, u is independent of time and the model reduces to the *stationary* reaction-diffusion equation

$$-(a(x)u'(x))' = f(x) \tag{53.7}$$

in the case of pure diffusion and

$$-(a(x)u'(x))' + (b(x)u(x))' = f(x) \tag{53.8}$$

in case there is convection as well. For these problems, we only need to specify boundary conditions. For example, we consider the *two-point boundary value problem*: find the function $u(x)$ satisfying

$$\begin{cases} -(au')' = f & \text{in } (0,1), \\ u(0) = u(1) = 0 \end{cases} \tag{53.9}$$

and when there is convection: find $u(x)$ such that

$$\begin{cases} -(au')' + (bu)' = f & \text{in } (0,1), \\ u(0,t) = u(1,t) = 0. \end{cases} \tag{53.10}$$

53.4 The Finite Element Method

We begin the discussion of discretization by studying the simplest model above, namely the two-point boundary value problem for the stationary reaction-diffusion model (53.9).

We can express the solution $u(x)$ of (53.9) analytically in terms of data by integrating twice (setting $w = au'$)

$$u(x) = \int_0^x \frac{w(y)}{a(y)}\,dy + \alpha_1, \quad w(y) = -\int_0^y f(z)\,dz + \alpha_2,$$

where the constants α_1 and α_2 are chosen so that $u(0) = u(1) = 0$. We can use this solution formula to compute the value of the solution $u(x)$ for any given $x \in (0,1)$ by evaluating the integrals analytically or numerically using quadrature. However, this is very time consuming if we want the solution at many points in $[0,1]$. This motivates consideration of an alternative way of computing the solution $u(x)$ using the *Finite Element Method* (FEM), which is a general method for solving differential equations numerically. FEM is based on rewriting the differential equation in *variational form* and seeking an approximate solution as a piecewise polynomial.

Note that we do not use the solution by integration outlined above, one important consequence of that procedure is that u is "twice as differentiable" as the data f, since we integrate twice to get from f to u.

We present FEM for (53.9) based on continuous piecewise linear approximation. We let $\mathcal{T}_h : 0 = x_0 < x_1 < ... < x_{M+1} = 1$, be a *partition* (or *triangulation*) of $I = (0,1)$ into sub-intervals $I_j = (x_{j-1}, x_j)$ of length $h_j = x_j - x_{j-1}$. We look for an approximate solution in the set V_h of continuous piecewise linear functions $v(x)$ on $\mathcal{T}_h$ such that $v(0) = 0$ and $v(1) = 0$. We show an example of such a function in Fig. 53.2. In Chapter 52, we saw

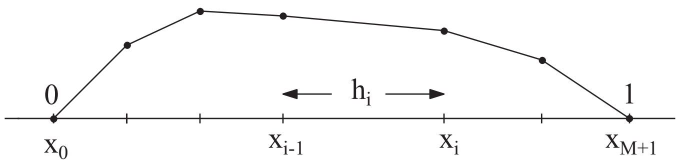

FIGURE 53.2. A continuous piecewise linear function in V_h.

that V_h is a finite dimensional vector space of dimension M with a basis consisting of the hat functions $\{\varphi_j\}_{j=1}^M$ illustrated in Fig. 52.5, associated with the *interior nodes* $x_1, \cdots, x_M$. The coordinates of a function v in V_h in this basis are the values $v(x_j)$ at the interior nodes since a function $v \in V_h$ can be written

$$v(x) = \sum_{j=1}^{M} v(x_j)\varphi_j(x).$$

Note that because $v \in V_h$ is zero at 0 and 1, we do not include φ_0 and φ_{M+1} in the set of basis functions for V_h.

The finite element method is based on restating the differential equation $-(au')' = f$ in an average or *variational* form

$$-\int_0^1 (au')'v\,dx = \int_0^1 fv\,dx, \tag{53.11}$$

where the function v varies over an appropriate set of *test functions*. The variational form results from multiplying the differential equation $-(au')' = f$ by the test function $v(x)$ and integrating over the interval $(0,1)$. The variational formulation says that the residual $-(au')' - f$ of the true solution is orthogonal to all test functions v with respect to the $L_2(0,1)$ scalar product.

The basic idea of FEM is to compute an approximate solution $U \in V_h$ that satisfies (53.11) for a restricted set of test functions. This approach to computing an approximate solution is known as the *Galerkin method* in memory of the Russian engineer and scientist Galerkin (1871-1945), see Fig. 53.3. He invented his method while imprisoned for anti-Tsarist activities during 1906-7. We call the set V_h, where we seek the FEM-solution U, the *trial space* and we call the space of test functions the *test space*. In the present case of homogeneous Dirichlet boundary conditions, we usually choose the test space to be equal to V_h. Consequently, the dimensions of the trial and test spaces are equal, which is necessary for the existence and uniqueness of the approximate solution U.

FIGURE 53.3. Boris Galerkin, inventor of the Finite Element Method

However since the functions in V_h do not have second derivatives, we can not simply plug a potential approximate solution U in V_h directly into (53.11). To get around this difficulty, we use integration by parts to move one derivative from $(au')'$ onto v, noting that functions in V_h are piecewise differentiable. Assuming v is differentiable and $v(0) = v(1) = 0$:

$$\begin{aligned} -\int_0^1 (au')'v\,dx &= -a(1)u'(1)v(1) + a(0)u'(0)v(0) + \int_0^1 au'v'\,dx \\ &= \int_0^1 au'v'\,dx. \end{aligned}$$

This leads to the *continuous Galerkin finite element method of order* 1 (*cG(1)-method*) for (53.9): compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h. \tag{53.12}$$

We note that the derivatives U' and v' of the functions U and $v \in V_h$ are piecewise constant functions of the form depicted in Fig. 53.4 and are not defined at the nodes x_i. However, the value of an integral is independent of the value of the integrand at isolated points. Therefore, the integral (53.12) with integrand $aU'v'$ is uniquely defined as the sum of the integrals over the sub-intervals I_j.

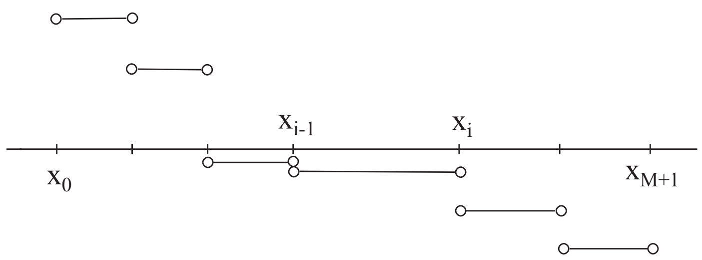

FIGURE 53.4. The derivative of the continuous piecewise linear function in Fig. 53.2.

53.4.1 Discretization of the stationary reaction-diffusion-convection problem

To solve (53.10) numerically let $0 = x_0 < x_1 < ... < x_{L+1} = 1$ be a partition of $(0, 1)$, and let V_h be the corresponding space of continuous piecewise linear functions $v(x)$ such that $v(0) = v(1) = 0$. The cG(1) FEM for (53.10) takes the form: compute $U \in V_h$ such that

$$\int_0^1 (aU')v' + (bU)'v)\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h.$$

53.5 The discrete system of equations

We have not yet proved that the set of equations (53.12) has a unique solution nor discussed what is involved in computing the solution U. This is an important issue considering we constructed the FEM precisely because the original problem is likely impossible to solve analytically.

We prove that the cG(1)-method (53.12) corresponds to a square linear system of equations for the unknown nodal values $\xi_j = U(x_j)$, $j = 1, ..., M$. We write U using the basis of hat functions as

$$U(x) = \sum_{j=1}^{M} \xi_j \varphi_j(x) = \sum_{j=1}^{M} U(x_j)\varphi_j(x).$$

Substituting into (53.12), we change the order of summation and integration to obtain

$$\sum_{j=1}^{M} \xi_j \int_0^1 a\varphi_j' v'\,dx = \int_0^1 fv\,dx, \tag{53.13}$$

for all $v \in V_h$. Now, it suffices to check (53.13) with v varying over the set of basis functions $\{\varphi_i\}_{i=1}^M$, since any function in V_h can be expressed as a linear combination of the basis functions. We are thus led to the $M \times M$ linear system of equations

$$\sum_{j=1}^{M} \xi_j \int_0^1 a\varphi_j'\varphi_i'\,dx = \int_0^1 f\varphi_i\,dx, \quad i = 1, ..., M, \tag{53.14}$$

for the unknown coefficients $\xi_1, ..., \xi_M$. We let $\xi = (\xi_1, ..., \xi_M)^\top$ denote the M-vector of unknown coefficients and define the $M \times M$ *stiffness matrix* $A = (a_{ij})$ with elements

$$a_{ij} = \int_0^1 a\varphi_j'\varphi_i'\,dx, \qquad i, j = 1, ..., M,$$

and the *load vector* $b = (b_i)$ with

$$b_i = \int_0^1 f\varphi_i\,dx, \qquad i = 1, ..., M.$$

These names originate from early applications of the finite element method in *structural mechanics* describing deformable structures like the body and wing of an aircraft or buildings. Using this notation, (53.14) is equivalent to the system of linear equations

$$A\xi = b. \tag{53.15}$$

In order to solve for the unknown vector ξ of nodal values of U, we first have to compute the stiffness matrix A and the load vector b. In the first instance, we assume that $a(x) = 1$ for $x \in [0, 1]$. We note that a_{ij} is zero unless $i = j - 1$, $i = j$, or $i = j + 1$ because otherwise either $\varphi_i(x)$ or $\varphi_j(x)$ is zero on each sub-interval occurring in the integration. We illustrate this in Fig. 53.5. We compute a_{ii} first. Using the definition of the hat function

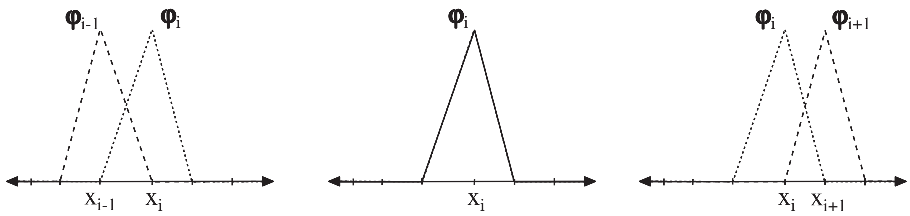

FIGURE 53.5. Three possibilities to obtain a non-zero element in the stiffness matrix.

φ_i,

$$\varphi_i(x) = \begin{cases} (x - x_{i-1})/h_i, & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)/h_{i+1}, & x_i \le x \le x_{i+1}, \\ 0, & \text{elsewhere}, \end{cases}$$

the integration breaks down into two integrals:

$$a_{ii} = \int_{x_{i-1}}^{x_i} \big(\frac{1}{h_i}\big)^2 \, dx + \int_{x_i}^{x_{i+1}} \big(\frac{-1}{h_{i+1}}\big)^2 \, dx = \frac{1}{h_i} + \frac{1}{h_{i+1}} \text{ for } i = 1, 2, ..., M,$$

since $\varphi_i' = 1/h_i$ on (x_{i-1}, x_i) and $\varphi_i' = -1/h_{i+1}$ on (x_i, x_{i+1}) and φ_i is zero on the other sub-intervals. Similarly,

$$a_{i\,i+1} = \int_{x_i}^{x_{i+1}} \frac{-1}{(h_{i+1})^2} \, dx = -\frac{1}{h_{i+1}} \text{ for } i = 1, 2, ..., M,$$

while $a_{i\,i-1} = -1/h_i$ for $i = 2, 3, ..., M$.

We compute the elements of the load vector of b in the same way to get

$$b_i = \int_{x_{i-1}}^{x_i} f(x) \frac{x - x_{i-1}}{h_i} dx + \int_{x_i}^{x_{i+1}} f(x) \frac{x_{i+1} - x}{h_{i+1}} dx, \quad i = 1, ..., M.$$

The matrix A is a *sparse* matrix in the sense that most of its entries are zero. In particular, A is a *banded* matrix with non-zero entries occurring only in the diagonal, super-diagonal and sub-diagonal positions. A is also called a tri-diagonal matrix. Moreover, A is a *symmetric* matrix since $\int_0^1 \varphi_i' \varphi_j' \, dx = \int_0^1 \varphi_j' \varphi_i' \, dx$. Finally, A is *positive-definite* in the sense that

$$\eta^\top A \eta = \sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j > 0,$$

unless $\eta_i = 0$ for $i = 1, ..., M$. This follows by noting that if $v(x) = \sum_{j=1}^{M} \eta_j \varphi_j(x)$ then by reordering the summation (check!)

$$\begin{aligned} \sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j &= \sum_{i,j=1}^{M} \eta_i \int_0^1 a \varphi_j' \varphi_i' \, dx \, \eta_j \\ &= \int_0^1 a \sum_{j=1}^{M} \eta_j \varphi_j' \sum_{i=1}^{M} \eta_i \varphi_i' \, dx = \int_0^1 a v'(x) v'(x) \, dx > 0 \end{aligned}$$

unless $v'(x) = 0$ for all $x \in [0, 1]$, that is $v(x) = 0$ for $x \in [0, 1]$, since $v(0) = 0$, that is $\eta_i = 0$ for $i = 1, ..., M$. This implies that A is invertible, so that (53.15) has a unique solution for all data b.

We sum up: the stiffness matrix A is sparse, symmetric and positive definite, and thus in particular the system $A\xi = b$ has a unique solution for all b.

We expect the accuracy of the approximate solution to increase as M increases since the work involved in solving for U increases. Systems of dimension $10^2 - 10^3$ in one space dimension and up to 10^6 in two or three space dimensions are common. An important issue is the efficient numerical solution of the system $A\xi = b$.

53.6 Discretizing problems with different boundary conditions

We consider briefly the discretization of the two-point boundary value problem $-(au')' = f$ in $(0, 1)$ with the different boundary conditions.

53.6.1 Non-homogeneous Dirichlet boundary conditions

We begin with the boundary conditions $u(0) = u_0$ and $u(1) = u_1$, where u_0 and u_1 are given boundary values, where the conditions are non-homogeneous if $u_0 u_1 \neq 0$. In this situation, we compute an approximate solution in the trial space V_h of continuous piecewise linear functions $v(x)$ on a partition $\mathcal{T}_h : 0 = x_0 < x_1 < ... < x_{M+1} = 1$, satisfying the boundary conditions $v(0) = u_0$, $v(1) = u_1$, and we let the test functions vary over the space V_h^0 of continuous piecewise linear functions $v(x)$ satisfying the homogeneous boundary conditions $v(0) = v(1) = 0$. The trial and test spaces are different in this case, but we note that they have equal dimension (equal to the number M of internal nodes). Multiplying by a test function and integrating by parts, we are led to the following method: compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h^0. \tag{53.16}$$

As above this leads to a symmetric positive definite system of equations in the internal unknown nodal values $U(x_1), ..., U(x_M)$.

53.6.2 Neumann boundary conditions

We now consider the problem

$$\begin{cases} -(au')' = f, & \text{in } (0,1), \\ u(0) = 0, \;\; a(1)u'(1) = g_1, \end{cases} \tag{53.17}$$

with a non-homogeneous Neumann boundary condition at $x = 1$, which in the case of modeling heat in a wire, corresponds to prescribing the heat flux $a(1)u'(1)$ at $x = 1$ to be g_1.

To derive a variational formulation of this problem, we multiply the differential equation $-(au')' = f$ by a test function v and integrate by parts to get

$$\int_0^1 fv\,dx = -\int_0^1 (au')'v\,dx = \int_0^1 au'v'\,dx - a(1)u'(1)v(1) + a(0)u'(0)v(0).$$

Now $a(1)u'(1) = g_1$ is specified but $a(0)u'(0)$ is unknown. So it is convenient to assume that v satisfies the homogeneous Dirichlet condition $v(0) = 0$.

Correspondingly, we define V_h to be the space of continuous functions v that are piecewise linear on a partition $\mathcal{T}_h$ of $(0,1)$ satisfying $v(0) = 0$. Replacing $a(1)u'(1)$ by g_1, we are led to the following FEM for (53.17): compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx = \int_0^1 fv\,dx + g_1 v(1) \quad \text{for all } v \in V_h. \tag{53.18}$$

We substitute $U(x) = \sum_{i=1}^{M+1} \xi_i \varphi_i(x)$, noting that the value $\xi_{M+1} = U(x_{M+1})$ at the node x_{M+1} is now undetermined, into (53.18) and choose $v = \varphi_1, \cdots, \varphi_{M+1}$ to get a $(M+1) \times (M+1)$ system of equations for ξ. We show the form of the resulting stiffness matrix with $a = 1$ and load vector in Fig. 53.6. Note that the last equation

$$\left(\begin{array}{c|c} A & \begin{matrix} 0 \\ \vdots \\ 0 \\ -h_{M+1}^{-1} \end{matrix} \\ \hline 0 \;\cdots\; 0 \;\; -h_{M+1}^{-1} & h_{M+1}^{-1} \end{array}\right) \qquad \left(\begin{array}{c} b \\ \hline b_{M+1}+g_1 \end{array}\right)$$

FIGURE 53.6. The stiffness matrix and load vector computed from (53.18) in the case that $a \equiv 1$. A and b are the stiffness matrix and load vector previously obtained in the problem with homogeneous Dirichlet boundary conditions and $b_{M+1} = \int_0^1 f\varphi_{M+1}\,dx$.

$$\frac{U(x_{M+1}) - U(x_M)}{h_{M+1}} = b_{M+1} + g_1$$

is a discrete analog of the boundary condition $u'(1) = g_1$ since $b_{M+1} \approx \frac{h_{M+1}}{2} f(1)$.

To conclude, a Neumann boundary condition, unlike a Dirichlet condition, is not explicitly enforced in the trial space. Instead, the Neumann condition is automatically satisfied as a consequence of the variational formulation by letting the test functions vary freely at the corresponding boundary point. In the case of Neumann boundary conditions, we thus simply can "forget" the boundary conditions in the definition of the trial space V_h and let the test space coincide with V_h. A Dirichlet boundary condition is called an *essential* boundary condition and a Neumann condition is called a *natural* boundary condition. An essential boundary condition is imposed explicitly in the definition of the trial space, i.e. it is a *strongly imposed* boundary condition, and the test space satisfy the corresponding homogeneous boundary condition. A natural boundary condition is not imposed in the trial space and becomes automatically satisfied through the variational formulation by letting the test functions vary freely at the corresponding boundary point.

53.6.3 Robin boundary conditions

A natural generalization of Neumann conditions for the problem $-(au')' = f$ in $(0,1)$ are called *Robin* boundary conditions. These take the form

$$-a(0)u'(0) = \gamma(0)(u_0 - u(0)), \quad a(1)u'(1) = \gamma(1)(u_1 - u(1)). \qquad (53.19)$$

In the case of modeling heat in a wire, $\gamma(0)$ and $\gamma(1)$ are given (non-negative) boundary heat conductivities and u_0 and u_1 are given "outside temperatures". The Robin boundary condition at $x = 0$ states that the heat flux $-a(0)u'(0)$ is proportional to the temperature difference $u_0 - u(0)$ between the outside and inside temperature. If $u_0 > u(0)$ then heat will flow from outside to inside and if $u_0 < u(0)$ then heat will flow from inside out.

EXAMPLE 53.3. We may experience this kind of boundary condition with $\gamma(0)$ quite large in a poorly insulated house on a cold winter day. The size of the boundary heat conductivity γ is an important issue in the real estate business in the north of Sweden.

When $\gamma = 0$, (53.19) reduces to a homogeneous Neumann boundary condition. Conversely, letting γ tend to infinity, the Robin boundary condition $-a(0)u'(0) = \gamma(0)(u_0 - u(0))$ approaches the Dirichlet boundary condition $u(0) = u_0$.

Robin boundary conditions are natural boundary conditions like Neumann conditions, Therefore, we let V_h be the space of continuous piecewise linear functions on a partition of $(0,1)$ without any boundary conditions imposed. Multiplying the equation $-(au')' = f$ by a function $v \in V_h$ and integrating by parts, we get

$$\int_0^1 fv\,dx = -\int_0^1 (au')'v\,dx = \int_0^1 au'v'\,dx - a(1)u'(1)v(1) + a(0)u'(0)v(0).$$

Replacing $a(0)u'(0)$ and $a(1)u'(1)$ using the Robin boundary conditions, we get

$$\int_0^1 fv\,dx = \int_0^1 au'v'\,dx + \gamma(1)(u(1) - u_1)v(1) + \gamma(0)(u(0) - u_0)v(0).$$

Collecting data on the right hand side, we are led to the following cG(1) method: compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx + \gamma(0)u(0)v(0) + \gamma(1)u(1)v(1)$$
$$= \int_0^1 fv\,dx + \gamma(0)u_0v(0) + \gamma(1)u_1v(1)$$

for all $v \in V_h$.

An even more general Robin boundary condition has the form $-a(0)u'(0) = \gamma(0)(u_0 - u(0)) + g_0$, where g_0 is a given heat flux. This Robin boundary condition thus includes Neumann boundary conditions ($\gamma = 0$) and Dirichlet boundary conditions (letting $\gamma \to \infty$). The implementation of a Robin boundary conditions is facilitated by the fact that the trial and test space are the same.

53.7 Error estimates and adaptive error control

When conducting scientific experiments in a laboratory or building a suspension bridge, for example, there is always a lot of worry about the errors in the process. In fact, if we were to summarize the philosophy behind the scientific revolution, a main component would be the modern emphasis on the quantitative analysis of error in measurements during experiments and the reporting of the errors along with the results. The same issue comes up in computational mathematical modeling: whenever we make a computation on a practical problem, we must be concerned with the accuracy of the results and the related issue of how to compute efficiently. These issues naturally fit into a wider framework which also addresses how well the differential equation models the underlying physical situation and what effect errors in data and the model have on the conclusions we can draw from the results.

We address these issues by deriving two kinds of error estimates for the error $u - U$ of the finite element approximation. First we prove an *a priori* error estimate which shows that the Galerkin finite element method for (53.9) produces the best possible approximation in V_h of the solution u in a certain sense. If u has continuous second derivatives, then we know that V_h contains good approximations of u, for example the piecewise linear interpolant. So the a priori estimate implies that the error of the finite element approximation can be made arbitrarily small by refining the mesh provided that the solution u is sufficiently smooth to allow the interpolation error to go to zero as the mesh is refined. This kind of result is called an *a priori* error estimate because the error bound does not depend on the approximate solution to be computed. One the other hand, it does requires knowledge about the derivatives of the (unknown) exact solution.

After that, we prove an *a posteriori* error bound that bounds the error of the finite element approximation in terms of its residual error. This error bound can be evaluated once the finite element solution has been computed and used to estimate the error. Through the a posteriori error estimate, it is possible to estimate and adaptively control the finite element error to a desired tolerance level by suitably refining the mesh.

To measure the size of the error $e = u - U$, we shall use the *weighted* L_2 *norm*

$$\|w\|_a = \left(\int_0^1 a\, w^2\, dx\right)^{1/2},$$

with *weight* a. More precisely we shall estimate the quantity

$$\|(u-U)'\|_a$$

which we refer to as the *energy norm* of the error $u - U$.

We will use the following variations of Cauchy's inequality with the weight a present:

$$\left|\int_0^1 av'w'\, dx\right| \le \|v'\|_a \|w'\|_a \quad \text{and} \quad \left|\int_0^1 vw\, dx\right| \le \|v\|_a \|w\|_{a^{-1}}. \tag{53.20}$$

53.7.1 An a priori error estimate

We shall prove that the finite element approximation $U \in V_h$ is the best approximation of u in V_h with respect to the energy norm. This is a consequence of the *Galerkin orthogonality* built into the finite element method expressed by

$$\int_0^1 a(u-U)'v'\, dx = 0 \quad \text{for all } v \in V_h \tag{53.21}$$

which results from subtracting (53.12) from (53.11) (integrated by parts) with $v \in V_h$. This is analogous to the best approximation property of the L_2 projection studied in the Chapter Piecewise linear approximation.

We have for any $v \in V_h$,

$$\begin{aligned}
\|(u-U)'\|_a^2 &= \int_0^1 a(u-U)'(u-U)'\, dx \\
&= \int_0^1 a(u-U)'(u-v)'\, dx + \int_0^1 a(u-U)'(v-U)'\, dx \\
&= \int_0^1 a(u-U)'(u-v)'\, dx,
\end{aligned}$$

where the last line follows because $v - U \in V_h$. Estimating using Cauchy's inequality, we get

$$\|(u-U)'\|_a^2 \le \|(u-U)'\|_a \|(u-v)'\|_a,$$

so that

$$\|(u - U)'\|_a \le \|(u - v)'\|_a \quad \text{for all } v \in V_h.$$

This is the best approximation property of U. We now choose in particular $v = \pi_h u$, where $\pi_h u \in V_h$ is the nodal interpolant of u, and use the following weighted analog of (52.11)

$$\|(u - \pi_h u)'\|_a \le C_i \|hu''\|_a,$$

where C_i is an interpolation constant that depends only on (the variation of) a. We then obtain the following error estimate.

Theorem 53.1 *The finite element approximation U satisfies $\|(u-U)'\|_a \le \|(u - v)'\|_a$ for all $v \in V_h$. In particular, there is a constant C_i depending only on a such that*

$$\|u' - U'\|_a \le C_i \|hu''\|_a.$$

This energy norm estimate says that the derivative of the error of the finite element approximation converges to zero at a first order rate in the mesh size h. By integration it follows that the error itself, say pointwise or in the L_2 norm, aslo tends to zero. One can also prove a more precise bound for the error $u - U$ itself that is second order in the mesh size h.

53.7.2 An a posteriori error estimate

We shall now estimate the energy norm error $\|u' - U'\|_a$ in terms of the residual $R(U) = (aU')' + f$ of the finite element solution U on each subinterval. The residual measures how well U solves the differential equation and it is completely computable once U has been computed.

We start by using the variational form of (53.11) with $v = e = u - U$ to find an expression for $\|u - U\|_a^2$:

$$\begin{aligned}\|e'\|_a^2 &= \int_0^1 ae'e'\,dx = \int_0^1 au'e'\,dx - \int_0^1 aU'e'\,dx \\ &= \int_0^1 fe\,dx - \int_0^1 aU'e'\,dx.\end{aligned}$$

We then use (53.12), with $v = \pi_h e$ denoting the nodal interpolant of e in V_h, to obtain

$$\begin{aligned}\|e'\|_a^2 &= \int_0^1 f\,(e - \pi_h e)\,dx - \int_0^1 aU'(e - \pi_h e)'\,dx \\ &= \int_0^1 f\,(e - \pi_h e)\,dx - \sum_{j=1}^{M+1} \int_{I_j} aU'(e - \pi_h e)'\,dx.\end{aligned}$$

Now, we integrate by parts over each sub-interval I_j in the last term and use the fact that all the boundary terms disappear because $(e - \pi_h e)(x_j) = 0$ to get the *error representation formula*

$$\|e'\|_a^2 = \int_0^1 R(U)(e - \pi_h e)\, dx, \tag{53.22}$$

where the residual error $R(U)$ is the discontinuous function defined on $(0,1)$ by

$$R(U) = f + (aU')' \quad \text{on each sub-interval } I_j.$$

From the weighted Cauchy inequality (53.20) (inserting factors h and h^{-1}), we get

$$\|e'\|_a^2 \leq \|hR(U)\|_{a^{-1}} \|h^{-1}(e - \pi_h e)\|_a.$$

One can prove the following analog of the second estimate of (52.11)

$$\|h^{-1}(e - \pi_h e)\|_a \leq C_i \|e'\|_a,$$

where C_i is an interpolation constant depending on a, and we notice the appearance of the factor h^{-1} on the left hand side. This proves the basic a posteriori error estimate:

Theorem 53.2 *There is an interpolation constant C_i depending only on a such that the finite element approximation U satisfies*

$$\|u' - U'\|_a \leq C_i \|hR(U)\|_{a^{-1}}. \tag{53.23}$$

53.7.3 Adaptive error control

Since the a posteriori error estimate (53.23) indicates the size of the error of an approximation on a given mesh in terms of computable information, it is natural to try to use this information to compute an accurate approximation. This is the basis of *adaptive error control.*

The computational problem that arises once a two-point boundary value problem is specified is to find a mesh such that the finite element approximation achieves a given level of accuracy, or in other words, such that the error of the approximation is bounded by an *error tolerance* TOL. In practice, we are also concerned with efficiency, which means that we want to determine a mesh with the fewest number of elements that yields an approximation with the desired accuracy. We try to reach this optimal mesh by starting with a coarse mesh and successively refining based on the size of the a posteriori error estimate. By starting with a coarse mesh, we try to keep the number of elements as small as possible.

More precisely, we choose an initial mesh $\mathcal{T}_h$, compute the corresponding cG(1) approximation U, and then check whether or not

$$C_i \|hR(U)\|_{a^{-1}} \leq \text{TOL}.$$

This is the *stopping criterion*, which guarantees that $\|u' - U'\|_a \leq \text{TOL}$ by (53.23). Therefore when the stopping criterion is satisfied, U is sufficiently accurate. If the stopping criterion is not satisfied, we try to construct a new mesh $\mathcal{T}_{\tilde{h}}$ of mesh size $\tilde{h}$ with as few elements as possible such that

$$C_i \|\tilde{h}R(U)\|_{a^{-1}} = \text{TOL}.$$

This is the *mesh modification criterion* from which the new mesh size $\tilde{h}$ is computed based on the size of the residual error $R(U)$ of the approximation on the old mesh. In order to minimize the number of mesh points, it turns out that the mesh size should be chosen to *equidistribute* the residual error in the sense that the contribution from each element to the integral giving the total residual error is roughly the same. In practice, this means that elements with large residual errors are refined, while elements in intervals where the residual error is small are combined together to form bigger elements.

We repeat the adaptive cycle of mesh modification followed by solution on the new mesh until the stopping criterion is satisfied. By the a priori error estimate, we know that if u'' is bounded then the error tends to zero as the mesh is refined. Hence, the stopping criterion will be satisfied eventually. In practice, the adaptive error control rarely requires more than a few iterations.

53.8 Discretization of time-dependent reaction-diffusion-convection problems

We now return to original time dependent problem (53.6).

To solve (53.6) numerically, we apply the cG(1) method for time discretization and the cG(1) FEM for discretization in space. More precisely, let $0 = x_0 < x_1 < \ldots < x_{L+1} = 1$ be a partition of $(0,1)$, and let V_h be the corresponding space of continuous piecewise linear functions $v(x)$ such that $v(0) = v(1) = 0$. Let $0 = t_0 < t_1 < t_2 < \ldots < t_N = T$ be a sequence of discrete time levels with corresponding time intervals $I_n = (t_{n-1}, t_n)$ and time steps $k_n = t_n - t_{n-1}$, for $n = 1, \ldots, N$. We look for a numerical solution $U(x,t)$ that is linear in t on each time interval I_n. For $n = 1, \ldots, N$, we compute $U^n \in V_h$ such that for all $v \in V_h$,

$$\begin{aligned}
&\int_{I_n}\int_0^1 \dot{U} v \, dx \, dt + \int_{I_n}\int_0^1 (aU')v' + (bU)'v) \, dx \, dt \\
&= \int_{I_n}\int_0^1 f v \, dx \, dt + \int_{I_n} (g(0,t)v(0) + g(1,t)v(1)) \, dt,
\end{aligned} \tag{53.24}$$

where $U(t_n, x) = U^n(x)$ denotes the *time nodal value* for $n = 1, 2, \ldots, N$ and $U^0 = u_0$, assuming that $u_0 \in V_h$. Since U is linear on each time interval, it is determined completely once we have computed its nodal values.

Arguing as above using the expansion in terms of the basis functions for V_h leads to a sequence of systems of equations for $n = 1, ..., N$,

$$MU^n + k_n A_n U^n = MU^{n-1} + k_n b^n, \tag{53.25}$$

where M is the mass matrix corresponding to V_h and A_n is a stiffness matrix related to time interval I_n. Solving this system successively for $n = 1, 2, ..., N$, we obtain an approximate solution U of (53.10).

53.9 Non-linear reaction-diffusion-convection problems

In many situations, the coefficients or data depend on the solution u, which leads to a nonlinear problem. For example if f depends on u, we get a problem of the form

$$\begin{cases} \dot{u} - (au')' + (bu)' = f(u) & \text{in } (0,1) \times (0,T), \\ u(0,t) = u(1,t) = 0, & \text{for } t \in (0,T), \\ u(x,0) = u_0(x) & \text{for } x \in (0,1). \end{cases} \tag{53.26}$$

Discretization as above eventually yields a discrete system of the form

$$MU^n + k_n A_n U^n = MU^{n-1} + k_n b^n(U^n), \tag{53.27}$$

where b^n depends on U^n. This nonlinear system may be solved by fixed point iteration or Newton's method.

We conclude this section by presenting some examples of systems of non-linear reaction-diffusion-convection problems arising in physics, chemistry and biology. These systems may be solved numerically by a direct extension of the cG(1) method in space and time presented above. In all examples, a and the α_i are a positive constants.

EXAMPLE 53.4. *The bistable equation for ferro-magnetism*

$$\dot{u} - au'' = u - u^3, \tag{53.28}$$

with a small.

EXAMPLE 53.5. *Model of a superconductivity of a fluid*

$$\begin{aligned} \dot{u}_1 - au_1'' &= (1 - |u|^2)u_1, \\ \dot{u}_2 - au_2'' &= (1 - |u|^2)u_2. \end{aligned} \tag{53.29}$$

EXAMPLE 53.6. *Model of flame propagation*

$$\begin{aligned} \dot{u}_1 - au_1'' &= -u_1 e^{-\alpha_1/u_2}, \\ \dot{u}_2 - au_2'' &= \alpha_2 u_1 e^{-\alpha_1/u_2}. \end{aligned} \tag{53.30}$$

EXAMPLE 53.7. *Field-Noyes equations for chemical reactions*

$$\begin{aligned}\dot{u}_1 - au_1'' &= \alpha_1(u_2 - u_1u_3 + u_1 - \alpha_2 u_1^2),\\ \dot{u}_2 - au_2'' &= \alpha^{-1}(\alpha_3 u_3 - u_2 - u_1u_2),\\ \dot{u}_2 - au_2'' &= \alpha_4(u_1 - u_3).\end{aligned} \tag{53.31}$$

EXAMPLE 53.8. *Spread of rabies in foxes*

$$\begin{aligned}\dot{u}_1 - au_1'' &= \alpha_1(1 - u_1 - u_2 - u_3) - u_3u_1,\\ \dot{u}_2 - au_2'' &= u_3u_1 - (\alpha_2 + \alpha_3 + \alpha_1u_1 + \alpha_1u_1 + \alpha_1u_3)u_2,\\ \dot{u}_2 - au_2'' &= \alpha_2u_2 - (\alpha_4 + \alpha_1u_1 + \alpha_1u_1 + \alpha_1u_3)u_3,\end{aligned} \tag{53.32}$$

where $\alpha_4 < (1 + (\alpha_3 + \alpha_1)/\alpha_2)^{-1} - \alpha_1$.

EXAMPLE 53.9. *Interaction of two species*

$$\begin{aligned}\dot{u}_1 - au_1'' &= u_1 M(u_1, u_2),\\ \dot{u}_2 - au_2'' &= u_2 N(u_1, u_2),\end{aligned} \tag{53.33}$$

where $M(u_1, u_2)$ and $N(u_1, u_2)$ are given functions describing various situations such as (i) predator-prey ($M_{u_2} < 0$, $N_{u_1} > 0$) (ii) competing species ($M_{u_2} < 0$, $N_{u_1} < 0$) and (iii) symbiosis ($M_{u_2} > 0$, $N_{u_1} > 0$).

EXAMPLE 53.10. *Morphogenesis of patterns (zebra or tiger)*

$$\begin{aligned}\dot{u}_1 - au_1'' &= -u_1u_2^2 + \alpha_1(1 - u_1)\\ \dot{u}_2 - au_2'' &= u_1u_2^2 - (\alpha_1 + \alpha_2)u_2.\end{aligned} \tag{53.34}$$

EXAMPLE 53.11. *Fitz-Hugh-Nagumo model for transmission of axons*

$$\begin{aligned}\dot{u}_1 - au_1'' &= -u_1(u_1 - \alpha_1)(u_1 - 1) - u_2\\ \dot{u}_2 - au_2'' &= \alpha_2u_1 - \alpha_3u_2,\end{aligned} \tag{53.35}$$

$0 < \alpha_1 < 1$.

Chapter 53 Problems

53.1. Compute the stiffness matrix and load vector for the cG(1) method on a uniform partition for (53.9) with $a(x) = 1 + x$ and $f(x) = \sin(x)$. Use quadrature if exact integration is inconvenient.

53.2. Formulate the cG(1) method for the problem $-(au')' + cu = f$ in $(0, 1)$, $u(0) = u(1) = 0$, where $a(x)$ and $c(x)$ are positive coefficients. Compute the corresponding stiffness matrix when $a = c = 1$, assuming a uniform partition. Is the stiffness matrix still symmetric, positive-definite, and tridiagonal?

53.3. Determine the resulting system of equations corresponding to the cG(1) method (53.16) with non-homogeneous Dirichlet boundary conditions.

53.4. Prove a priori and a posteriori error estimates for cG(1) for $-(au')' = f$ in $(0,1)$ with Robin boundary conditions (a positive).

53.5. Prove a priori and a posteriori error estimates for cG(1) for $-(au')'+cu = f$ in $(0,1)$ with Robin boundary conditions (a and c positive).

> The "classical" phase of my career was summed up in the book *The Large Scale Structure of Spacetime* which Ellis and I wrote in 1973. I would not advise readers of this book to consult that work for further information: it is highly technical, and quite unreadable. I hope that since then I have learned how to write in a manner that is easier to understand. (Stephen Hawking in *A Brief History of Time*)