# Chapter 3

# Pinning It Down: Boundary Value Problems

## 3.1 The Finite Difference Method for Linear Problems

In the previous section we considered initial value problems for ordinary differential equations. In many problems, however, there will be conditions on the solution given at more than one point. For a single first order equation $y' = f(x, y)$, data at one point completely determines the solution so that if conditions at more than one point are given, either higher order equations or systems of equations must be treated. Consider the second-order equation

$$v''(x) = f(x, v(x), v'(x)), \qquad 0 \leq x \leq 1. \tag{3.1.1}$$

Because it is second-order, such an equation requires two additional conditions, and the simplest possibility is to prescribe the solution values at the end-points:

$$v(0) = \alpha, \qquad v(1) = \beta. \tag{3.1.2}$$

Equations (3.1.1) and (3.1.2) define a *two-point boundary value* problem. We note that the restriction of the interval in (3.1.1) to $[0, 1]$ is no loss of generality since a problem on any finite interval may be converted to one on the interval $[0, 1]$ (Exercise 3.1.1). Alternatively, the following development is easily done directly on any finite interval (Exercise 3.1.6).

If the function $f$ of (3.1.1) is nonlinear in either $v(x)$ or $v'(x)$, the boundary value problem is *nonlinear*. Nonlinear boundary value problems are intrinsically more difficult to solve, and we will consider them in Chapter 5. In this

chapter we treat only linear problems, in which case (3.1.1) may be written in the form

$$v''(x) = b(x)v'(x) + c(x)v(x) + d(x), \qquad 0 \le x \le 1, \tag{3.1.3}$$

where $b$, $c$, and $d$ are given functions of $x$. The boundary conditions that we consider first will be (3.1.2); later, we shall treat other types of boundary conditions. Equations (3.1.3) and (3.1.2) define a linear two-point boundary-value problem for the unknown function $v$, and our task is to develop procedures to approximate the solution. We will assume that the problem has a unique solution that is at least two times continuously differentiable.

We first consider the special case of (3.1.3) in which $b(x) \equiv 0$, so that the equation is

$$v''(x) = c(x)v(x) + d(x), \qquad 0 \le x \le 1. \tag{3.1.4}$$

We will assume that $c(x) \ge 0$ for $0 \le x \le 1$; this is a sufficient condition for the problem (3.1.4), (3.1.2) to have a unique solution. To begin the numerical solution we divide the interval $[0, 1]$ into a number of equal subintervals of length $h$, as shown in Figure 3.1. As in Chapter 2, in Figure 3.1 the points $x_i$ are called the grid points, or nodes, and $h$ is the grid spacing; $x_0$ and $x_{n+1}$ are the *boundary* points, and $x_1, \ldots, x_n$ are the *interior* grid points.


Figure 3.1: *Grid Points*

We now need to approximate $v''(x)$ in (3.1.4), and we do this by *finite differences*. Let $x_i$ be any interior grid point and approximate the first derivatives at the points $x_i \pm \frac{h}{2}$ by

$$v'(x_i - \tfrac{h}{2}) \doteq \frac{[v(x_i) - v(x_{i-1})]}{h}, \qquad v'(x_i + \tfrac{h}{2}) \doteq \frac{[v(x_{i+1}) - v(x_i)]}{h}.$$

These relations are then used to approximate the second derivative:

$$v''(x_i) \doteq \frac{v'(x_i + \frac{h}{2}) - v'(x_i - \frac{h}{2})}{h} \doteq \frac{v(x_{i+1}) - 2v(x_i) + v(x_{i-1})}{h^2}. \tag{3.1.5}$$

If we now put this approximation into equation (3.1.4) and denote the functions $c$ and $d$ evaluated at $x_i$ by $c_i$ and $d_i$, we obtain

$$\frac{1}{h^2}[v(x_{i+1}) - 2v(x_i) + v(x_{i-1})] - c_i v(x_i) \doteq d_i, \quad i = l, \ldots, n. \tag{3.1.6}$$

What we have shown so far is that if we replace the second derivatives of the solution $v$ by finite difference approximations and put these approximations into the differential equation, we obtain the approximate relations (3.1.6) that the solution must satisfy. We now turn this procedure around. Suppose that we can find numbers $v_1, \ldots, v_n$ that satisfy the equations

$$\frac{1}{h^2}(v_{i+1} - 2v_i + v_{i-1}) - c_i v_i = d_i, \quad i = 1, \ldots, n, \tag{3.1.7}$$

or

$$-v_{i+1} + 2v_i - v_{i-1} + c_i h^2 v_i = -h^2 d_i, \quad i = 1, \ldots, n, \tag{3.1.8}$$

with $v_0 = \alpha$ and $v_{n+1} = \beta$. Then we can consider $v_1, \ldots, v_n$ to be approximations at the grid points $x_1, \ldots, x_n$ to the solution $v$ of the boundary-value problem (3.1.4), (3.1.2). We shall return shortly to the question of the accuracy of these approximations.

The equations (3.1.8) form a system of $n$ linear equations in the $n$ unknowns $v_1, \ldots, v_n$ and can be written in matrix-vector form as

$$\begin{bmatrix} 2+c_1h^2 & -1 & & & \\ -1 & 2+c_2h^2 & \ddots & & \\ & \ddots & \ddots & & \\ & & & -1 & \\ & & & -1 & 2+c_nh^2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} -h^2d_1 + \alpha \\ -h^2d_2 \\ \vdots \\ -h^2d_{n-1} \\ -h^2d_n + \beta \end{bmatrix}. \tag{3.1.9}$$

Thus, to obtain the approximate solution $v_1, \ldots, v_n$, we need to solve this system of linear equations. Techniques for this will be discussed in the next section.

The coefficient matrix of (3.1.9) in the case that the $c_i$ are all zero is

$$\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & & \\ & & & -1 & \\ & & & -1 & 2 \end{bmatrix}. \tag{3.1.10}$$

This is an important matrix which arises in many contexts, as we shall see. Matrices of the form (3.1.9) or (3.1.10) are called *tridiagonal* since only the three main diagonals of the matrix have non-zero elements. Tridiagonal matrices arise in a variety of applications in addition to the two-point boundary value problems of this chapter.

**Discretization Error**

We next consider the important question of the error in the approximations $v_1, \ldots, v_n$. Since the linear system (3.1.9) which determines these quantities will be solved numerically, the computed $v_i$ will be in error because of rounding; this will be discussed in more detail in Chapter 4. For the present we assume that the $v_i$ are computed with no rounding error so that $v_1, \ldots, v_n$ is the exact solution of the system (3.1.9). Let $v(x_i)$ again be the exact solution of the boundary-value problem at $x_i$. Then, analogous to the definition for initial-value problems in Chapter 2,

$$\max_{1 \le i \le n} |v_i - v(x_i)| \tag{3.1.11}$$

is the *(global) discretization error*. We now indicate how an analysis of the discretization error proceeds, and give results in a particular case. Again, we restrict our attention to the equation (3.1.4).

We first define the *local discretization error*, in a manner analogous to that for initial-value problems, by

$$L(x,h) = \frac{1}{h^2}[v(x+h) - 2v(x) + v(x-h)] - c(x)v(x) - d(x), \tag{3.1.12}$$

where $v$ is the exact solution of the differential equation (3.1.4) and $h$ is the grid spacing. By means of (3.1.4), we can replace $cv + d$ in (3.1.12) with $v''$ so that

$$L(x,h) = \frac{1}{h^2}[v(x+h) - 2v(x) + v(x-h)] - v''(x). \tag{3.1.13}$$

Thus the local discretization error is just the error in approximating $v''$. To estimate this error, we assume that $v$ is four times continuously differentiable and expand $v(x+h)$ and $v(x-h)$ in Taylor series. After we collect terms (the details of which are left to Exercise 3.1.2), we obtain

$$L(x,h) = \frac{1}{12}v^{(4)}(x)h^2 + 0(h^4) = 0(h^2). \tag{3.1.14}$$

The problem now is to relate this local discretization error to the global error (3.1.11). To do this we evaluate (3.1.12) at the grid points $x_i$, set $\sigma_i = L(x_i, h)$, and then subtract (3.1.7) from (3.1.12). Setting $e_i = v(x_i) - v_i$, this gives

$$\sigma_i = \frac{1}{h^2}[e_{i+1} - 2e_i + e_{i-1}] - c_i e_i, \quad i = 1, \cdots, n,$$

or

$$(2 + c_i h^2)e_i - e_{i+1} - e_{i-1} = -h^2\sigma_i, \quad i = l, \ldots, n, \tag{3.1.15}$$

where $e_0 = e_{n+1} = 0$. If $A$ is the coefficient matrix of (3.1.9) and $\mathbf{e}$ and $\boldsymbol{\sigma}$ are vectors with components $e_1, \ldots, e_n$ and $\sigma_1, \ldots, \sigma_n$, we can write (3.1.15) as

$$A\mathbf{e} = -h^2\boldsymbol{\sigma}, \tag{3.1.16}$$

or, assuming that $A^{-1}$ exists,

$$\mathbf{e} = -h^2 A^{-1}\boldsymbol{\sigma}. \tag{3.1.17}$$

This is the basic relationship between the global and local discretization errors. Note that the global discretization errors $e_1, \ldots, e_n$ and the approximate solutions $v_1, \ldots, v_n$ satisfy systems of equations with exactly the same coefficient matrix, but different right hand sides.

The problem now is to study the behavior of $A^{-1}$ as $h \to 0$. This is made more difficult by the fact that $n$, the order of $A$, tends to infinity as $h \to 0$. It is beyond the scope of this book to pursue this problem in any generality, but we can give a relatively simple analysis of the discretization error in the case where

$$c(x) \geq \gamma > 0, \qquad x \in [0, 1], \tag{3.1.18}$$

so that $c_i \geq \gamma$, $i = 1, \ldots, n$. If we set $e = \max |e_i|$ and $\sigma = \max |\sigma_i| = 0(h^2)$, we obtain from (3.1.15) and (3.1.18) that

$$(2 + \gamma h^2)|e_i| \leq 2e + h^2\sigma, \qquad i = 1, \ldots, n. \tag{3.1.19}$$

Since (3.1.19) holds for all $i$, we must have

$$(2 + \gamma h^2)e \leq 2e + h^2\sigma.$$

Therefore, since by (3.1.14) $\sigma = 0(h^2)$, we conclude that

$$e \leq \frac{\sigma}{\gamma} = 0(h^2), \tag{3.1.20}$$

which shows that the global discretization error is $0(h^2)$ provided that the local discretization is $0(h^2)$. It can be shown that the same result holds more generally, particularly for the important special case in which $c(x) \equiv 0$, but a more difficult analysis is required.

**More General Equations**

We next consider the more general equation (3.1.3), in which $v'$ is present. The standard centered difference approximation to $v'(x)$ is

$$v'(x) \doteq \frac{1}{2h}[v(x + h) - v(x - h)], \tag{3.1.21}$$

and it is easy to show (Exercise 3.1.3) that the error in this approximation is $0(h^2)$. If we replace $v'(x)$ in (3.1.3) at the grid points and proceed as before, the equations corresponding to (3.1.8) now become

$$\left(-1 - \frac{b_i h}{2}\right) v_{i-1} + (2 + c_i h^2)v_i + \left(-1 + \frac{b_i h}{2}\right) v_{i+1} = -h^2 d_i, \quad i = 1, \ldots, n,$$

or

$$r_i v_{i-1} + p_i v_i + q_i v_{i+1} = -h^2 d_i, \quad i = 1, \ldots, n,$$

where we have set

$$p_i = 2 + c_i h^2, \qquad q_i = -1 + \frac{b_i h}{2}, \qquad r_i = -1 - \frac{b_i h}{2}. \tag{3.1.22}$$

Then we can write these equations in matrix form as

$$\begin{bmatrix} p_1 & q_1 & & & \\ r_2 & p_2 & q_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & & q_{n-1} \\ & & & r_n & p_n \end{bmatrix} \begin{bmatrix} v_1 \\ \\ \vdots \\ \\ v_n \end{bmatrix} = -h^2 \begin{bmatrix} d_1 + r_1\alpha/h^2 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n + q_n\beta/h^2 \end{bmatrix}. \tag{3.1.23}$$

### Diagonal Dominance

One desirable property of the coefficient matrix of (3.1.23) is diagonal dominance. A general $n \times n$ matrix $A = (a_{ij})$ is *row diagonally dominant* if

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|, \quad i = 1, \ldots, n; \tag{3.1.24}$$

that is, the absolute value of the diagonal element of each row is at least as large as the sum of the absolute values of all the off-diagonal elements in that row. The matrix is *column diagonally dominant* if

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ji}|, \quad i = 1, \ldots, n. \tag{3.1.25}$$

Diagonal dominance is important for a number of reasons, one of which is that it is an approach to showing non-singularity of the matrix. Diagonal dominance by itself is not sufficient for this; for example, a $2 \times 2$ matrix with all elements equal to 1 is diagonally dominant but singular. However, the following strengthening of diagonal dominance guarantees non-singularity. The matrix $A$ is *strictly* row (column) diagonally dominant if strict inequality holds in (3.1.24) for all $i$ (or (3.1.25) for column). We then have the following result.

THEOREM 3.1.1. *If the $n \times n$ matrix $A$ is strictly row or column diagonally dominant, then it is non-singular.*

To prove this theorem, assume that $A$ is singular so that $A\mathbf{x} = 0$ for some non-zero $\mathbf{x}$ (see Theorem A.2.1 in Appendix 2). Let $|x_k| = \max\{|x_i| : i = 1, \ldots, n\}$. The $k$th equation of $A\mathbf{x} = 0$ is

$$a_{kk} x_k = -\sum_{j \neq k} a_{kj} x_j$$

so that

$$|a_{kk}||x_k| = |a_{kk}x_k| = |\sum_{j\neq k} a_{kj}x_j| \leq |x_k| \sum_{j\neq k} |a_{kj}|.$$

By assumption, $|x_k| > 0$ so it can be divided out and the resulting inequality contradicts strict row diagonal dominance. This contradiction shows that $A$ is non-singular. If $A$ is strictly column diagonally dominant, then $A^T$, the transpose of $A$, is strictly row diagonally dominant, so it is non-singular. Hence $A$ is non-singular since $\det A = \det A^T$. This completes the proof.

For the matrix of (3.1.23) to be row diagonally dominant we need that

$$|p_i| \geq |r_i| + |q_i|, \qquad i = 1, \ldots, n,$$

or, using (3.1.22),

$$|2 + c_i h^2| \geq |1 + \frac{b_i h}{2}| + |1 - \frac{b_i h}{2}|, \qquad i = 1, \ldots, n. \tag{3.1.26}$$

If we assume that $c_i \geq 0$, then (3.1.26) holds if $h$ is sufficiently small. In particular, if

$$|b_i h| \leq 2, \qquad i = 1, \ldots, n, \tag{3.1.27}$$

then the absolute values of the quantities on the right side of (3.1.26) are the quantities themselves, so that

$$|2 + c_i h^2| \geq 1 + \frac{b_i h}{2} + 1 - \frac{b_i h}{2} = 2.$$

The condition (3.1.27) on $h$, which also ensures column diagonal dominance, is a rather stringent one and can be avoided by using one-sided differences in place of the central difference (3.1.21) to approximate the first derivative. More precisely, we use the approximations

$$v'(x_i) \doteq \begin{cases} \frac{1}{h}(v_{i+1} - v_i) & \text{if} \quad b_i < 0, \\ \frac{1}{h}(v_i - v_{i-1}) & \text{if} \quad b_i \geq 0, \end{cases} \tag{3.1.28}$$

so that the direction of the one-sided difference is determined by the sign of $b_i$. Such differences are quite commonly used in fluid dynamics problems and in that context are called *upwind* (or *upstream*) differences. With (3.1.28) the $i$th row of the coefficient matrix (3.1.23) becomes

$$\begin{array}{llll} -1 & 2 + c_i h^2 - b_i h & -1 + b_i h, & b_i \leq 0, \\ -(1 + b_i h) & 2 + c_i h^2 + b_i h & -1, & b_i > 0, \end{array} \tag{3.1.29}$$

and it is easy to verify that row diagonal dominance holds, independent of the size of $h$, assuming again that $c_i \geq 0$. We note, however, that although the

centered difference approximation (3.1.21) is second-order accurate, the one-sided approximations (3.1.28) are only first-order accurate, and this increase in the discretization error must be weighed against the better properties of the coefficient matrix.

### Symmetry

Another desirable property of the coefficient matrix of (3.1.23) is symmetry. A matrix $A = (a_{ij})$ is *symmetric* if $A = A^T$; that is, if

$$a_{ij} = a_{ji}, \quad i, j = 1, \ldots, n. \tag{3.1.30}$$

The matrix of (3.1.9) is symmetric, but for the matrix of (3.1.23) to be symmetric we need that $q_i = r_{i+1}$, or using (3.1.22),

$$-1 + \tfrac{1}{2}b_i h = -1 - \tfrac{1}{2}b_{i+1}h, \quad i = 1, \ldots, n-1.$$

Clearly, these relations will not usually hold. In many situations, however, the differential equation is of the form

$$[a(x)v']' = d(x). \tag{3.1.31}$$

In this case we can obtain a symmetric coefficient matrix by "symmetric differencing," as described in the Supplementary Discussion.

### Other Boundary Conditions

The previous discussion has used the boundary conditions (3.1.2). In many problems, however, boundary conditions on the derivative rather than the function itself may be given, and we now consider the modifications that this requires.

Suppose, for example, that we have the boundary conditions

$$v'(0) = \alpha \qquad v'(1) = \beta \tag{3.1.32}$$

in place of (3.1.2); that is, we specify derivative values rather than function values. Consider now the difference equations (3.1.8). In the equation for $i = 1$, the value of $v_0$ is no longer known from the boundary condition at $x = 0$. Instead $v_0$ will be an additional unknown and we will need another equation that can be derived as follows. We approximate $v''$ at $x = 0$ by

$$v''(0) \doteq \frac{1}{h^2}[v_{-1} - 2v_0 + v_1], \tag{3.1.33}$$

using a grid point $-h$ outside the interval. Then by the boundary condition

$$\alpha = v'(0) \doteq \frac{1}{2h}[v_1 - v_{-1}] \tag{3.1.34}$$

we have $v_{-1} \doteq v_1 - 2\alpha h$, and we can use this to eliminate $v_{-1}$ in (3.1.33):

$$v''(0) \doteq \frac{1}{h^2}[2v_1 - 2v_0 - 2\alpha h]. \tag{3.1.35}$$

In general this approximation is only first-order accurate, but in the important special case that $\alpha = 0$ it is second-order (Exercise 3.1.7). We can now use (3.1.35) to obtain an additional equation and we have $n+1$ equations in the $n+1$ unknowns $v_0, \ldots, v_n$. If $v(1) = \beta$ is specified as before, then this is the system of equations to be solved. If $v'(1)$ is specified, as in (3.1.32), then another equation would arise from the approximation

$$v''(1) \doteq \frac{1}{h^2}[2v_n - 2v_{n+1} + 2\beta h], \tag{3.1.36}$$

and $v_{n+1}$ would be an additional unknown.

If only the function value is specified, as in (3.1.2), the boundary conditions are called *Dirichlet*, whereas if only the derivatives are specified, as in (3.1.32), the boundary conditions are called *Neumann*. More generally, mixed boundary conditions may be given as linear combinations of both function and derivative values at both end points:

$$\eta_1 v(0) + \eta_2 v'(0) = \alpha, \qquad \gamma_1 v(1) + \gamma_2 v'(1) = \beta.$$

In this case approximations analogous to those discussed previously would be used.

We return to the boundary conditions (3.1.32), and consider the differential equation (3.1.4). The approximations (3.1.35) and (3.1.36) give the two equations

$$(2 + c_0 h^2)v_0 - 2v_1 = 2\alpha h - h^2 d_0,$$
$$(2 + c_{n+1} h^2)v_{n+1} - 2v_n = 2\beta h - h^2 d_{n+1},$$

which are added to the system (3.1.8) to give $n+2$ equations in the $n+2$ unknowns $v_0, \ldots, v_{n+1}$. The coefficient matrix of this system is

$$A = \begin{bmatrix} 2 + c_0 h^2 & -2 & & & \\ -1 & 2 + c_1 h^2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & & -1 \\ & & & -2 & 2 + c_n h^2 \end{bmatrix}. \tag{3.1.37}$$

This matrix is no longer symmetric, but it can be easily symmetrized if desired (Exercise 3.1.8). If $c_i > 0$, $i = 0, \ldots, n+1$, Theorem 3.1.1 shows that $A$ is non-singular. But if the $c_i$ are all zero it is singular, since $A\mathbf{e} = 0$, where

$\mathbf{e} = (1, 1, \ldots, 1)^T$. This singularity of $A$ mirrors the non-uniqueness of solutions of the differential equation itself: if $c(x) \equiv 0$ in (3.1.4) and $v$ is a solution satisfying the boundary conditions (3.1.32), then $v + \gamma$ is also a solution for any constant $\gamma$.

Another type of boundary condition that leads to non-unique solutions is called *periodic*:

$$v(0) = v(1). \tag{3.1.38}$$

In this case $v_0$ and $v_{n+1}$ are both unknowns but since they are equal only one needs to be added to the system of equations. Again, we can use (3.1.33) as an additional equation. By (3.1.38), we can assume that the solution $v$ is extended periodically outside $[0, 1]$; in particular, we can take $v_{-1} = v_n$ in (3.1.33) so that it becomes

$$v''(0) \doteq \frac{1}{h^2}[v_n - 2v_0 + v_1]. \tag{3.1.39}$$

Similarly, in the approximation at $x_n$ we can take $v_{n+1} = v_0$:

$$v''(x_n) \doteq \frac{1}{h^2}[v_{n-1} - 2v_n + v_{n+1}] = \frac{1}{h^2}[v_{n-1} - 2v_n + v_0]. \tag{3.1.40}$$

There will then be $n + 1$ unknowns $v_0, \ldots, v_n$, and the coefficient matrix for the problem (3.1.4) can be obtained, using (3.1.39) and (3.1.40), in the form

$$A = \begin{bmatrix} 2 + c_0 h^2 & -1 & & -1 \\ -1 & 2 + c_1 h^2 & \ddots & \\ & \ddots & \ddots & -1 \\ -1 & & -1 & 2 + c_n h^2 \end{bmatrix}. \tag{3.1.41}$$

The tridiagonal structure has now been lost because of the outlying $-1$'s. As with (3.1.37), it follows from Theorem 3.1.1 that the matrix $A$ of (3.1.41) is non-singular if all $c_i$ are positive, but singular if all $c_i$ are zero. (As before, $A\mathbf{e} = 0$.) This singularity again reflects the differential equation, since if $c(x) \equiv 0$ in (3.1.4) and $v$ is a solution satisfying (3.1.38), then $v + \gamma$ is also a solution for any constant.

Whatever difference approximation or boundary conditions are involved, the basic computational problem is to solve the resulting system of linear equations. We will address this question in the next section for the particular systems arising from two-point boundary value problems. The next chapter will consider linear systems in more generality.

## Supplementary Discussion and References: 3.1

We have discussed in the text only the simplest differencing procedures for rather simple problems. As we saw, the central difference approximation (3.1.5) gives rise to discretization error proportional to $h^2$. It is frequently desirable to use more accurate methods. One approach is to use higher-order approximations; for example,

$$v''(x_i) \doteq \frac{1}{12h^2}(-v_{i-2} + 16v_{i-1} - 30v_i + 16v_{i-1} - v_{i+2}) \tag{3.1.42}$$

is a fourth-order approximation (the error is proportional to $h^4$), provided that $v$ is sufficiently differentiable. One difficulty with applying approximations of this type to two-point boundary-value problems occurs near the boundary. For example, if we apply the approximation at the first interior grid point $x_1$, it requires values of $v$ not only at $x_0$ but also at $x_{-1}$, which is outside the interval. However, second-order approximations can be applied at $x_1$ and $x_{n-1}$ and still retain fourth-order accuracy of the solution. This was proved by Shoosmith [1973] for various second order approximations, depending on (3.1.3), using techniques developed by Bramble and Hubbard [1964] for partial differential equations.

Another approach to obtaining higher-order approximations to the solution, while using only second-order approximations to the derivatives is Richardson extrapolation. This was discussed in Section 2.2 for initial value problems and applies also to boundary value problems. Let $x^*$ be some fixed point in the interval $[a, b]$ and let $v(x^*; h)$ denote the approximate solution at $x^*$ as a function of $h$. Assume that

$$v(x^*; h) = v(x^*) + c_2h^2 + c_3h^3 \cdots + c_ph^p + 0(h^{p+1}) \tag{3.1.43}$$

where $v(x^*)$ is the solution of the differential equation at $x^*$, and the $c_i$ are functions of $x^*$ and the solution. Note that the expansion (3.1.43) starts with $h^2$ since we are assuming that the method is second order accurate. If $v(x^*; \frac{h}{2})$ is the approximate solution obtained with step length $\frac{h}{2}$, then

$$\bar{v}(x^*; h) \equiv \frac{1}{3}[4v(x^*; \frac{h}{2}) - v(x^*; h)] = d_3h^3 + \cdots + 0(h^{p+1}),$$

so that this new approximation is third-order accurate. In certain cases, such as when the first derivative is absent in (3.1.3), the error expansion (3.1.43) will contain only even powers of $h$, and one application of the extrapolation principle will then give fourth-order accuracy and is particularly effective.

Still another approach to higher-order accuracy is *deferred correction.* In this procedure one computes an approximation to the local discretization error by means of the current approximate solution, and then uses this to obtain a new approximate solution with $0(h^4)$ accuracy. The process can then be

repeated to obtain still higher order approximations. For further discussion of this method, see, for example, Ascher et al. [1988].

We next give more details on "symmetric differencing" for the "self-adjoint" equation (3.1.31). As in the derivation of (3.1.5), we use the auxiliary grid points $x_i \pm \frac{h}{2}$ and approximate the outermost derivative of $(av')'$ at $x_i$ by

$$[a(x)v'(x)]'_i \doteq \frac{1}{h}(a_{i+\frac{1}{2}}v'_{i+\frac{1}{2}} - a_{i-\frac{1}{2}}v'_{i-\frac{1}{2}}), \tag{3.1.44}$$

where the subscripts indicate the grid point at which the evaluation is done. We approximate the first derivatives by the centered differences

$$v'_{i+\frac{1}{2}} \doteq \frac{1}{h}(v_{i+1} - v_i) \qquad v'_{i-\frac{1}{2}} \doteq \frac{1}{h}(v_i - v_{i-1})$$

and use these in (3.1.44) to obtain

$$\begin{aligned} [a(x)v'(x)]'_i &\doteq \frac{1}{h^2}[a_{i+\frac{1}{2}}(v_{i+1} - v_i) - a_{i-\frac{1}{2}}(v_i - v_{i-1})] \qquad (3.1.45) \\ &= \frac{1}{h^2}[a_{i-\frac{1}{2}}v_{i-1} - (a_{i+\frac{1}{2}} + a_{i-\frac{1}{2}})v_i + a_{i+\frac{1}{2}}v_{i+1}]. \end{aligned}$$

Therefore the system of difference equations corresponding to (3.1.31) is of the form $A\mathbf{v} = \mathbf{q}$ where

$$A = \begin{bmatrix} a_{3/2} + a_{1/2} & -a_{3/2} & & & \\ -a_{3/2} & a_{5/2} + a_{3/2} & -a_{5/2} & & \\ & \ddots & \ddots & \ddots & \\ & & & & -a_{n-1/2} \\ & & & -a_{n-1/2} & a_{n-1/2} + a_{n+1/2} \end{bmatrix}. \tag{3.1.46}$$

Clearly, this matrix is both symmetric and diagonally dominant. Note also that if $a(x) \equiv 1$, (3.1.46) reduces to the $(2, -1)$ matrix (3.1.10). If a term $c(x)v$ is also present in (3.1.31), this will just add $c_i h^2$ to the $i$th diagonal element of (3.1.46).

A matrix $A$ is *reducible* if there is a permutation matrix $P$ (see Section 4.3) so that $PAP^T$ has the form

$$PAP^T = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix},$$

where $A_1$ and $A_3$ are square submatrices. If no such permutation matrix can be found, then $A$ is *irreducible*. It can be shown that any tridiagonal matrix whose off-diagonal elements are non-zero is irreducible; in particular, the

matrix (3.1.10) is irreducible. A matrix that is irreducible and row diagonally dominant is *irreducibly diagonally dominant* if strict inequality holds in (3.1.24) for at least one $i$. The proof of Theorem 3.1.1 can be extended to show that irreducibly diagonally dominant matrices are non-singular; in particular, the matrix (3.1.10) is non-singular.

For further reading on two-point boundary value problems, see Ortega [1990] and the excellent books by Ascher et al. [1988] and Keller [1968].

## EXERCISES 3.1

**3.1.1.** Consider the boundary value problem $u''(z) = g(z, u(z), u'(z))$ with $u(a) = \alpha$, $u(b) = \beta$. By the change of variable $t = (b-a)x + a$, show that this problem is equivalent to (3.1.1) and (3.1.2) with $v(x) = u((b-a)x + a)$ and $f(x, v(x), v'(x)) = (b-a)^2 g((b-a)x + a, v(x), (b-a)^{-1}v'(x))$. Specialize this result to the linear problem (3.1.3).

**3.1.2.** Assume that the function $v(x)$ is suitably differentiable. By expanding $v(x+h)$ and $v(x-h)$ in Taylor series, verify that (3.1.14) holds.

**3.1.3** Assume that the function $v(x)$ is twice differentiable. Show that the error in the centered difference approximation (3.1.21) is proportional to $h^2$.

**3.1.4.** Consider the two-point boundary-value problem

$$v'' + 2xv' - x^2 v = x^2, v(0) = 1, v(1) = 0.$$

**a.** Let $h = \frac{1}{4}$ and explicitly write out the difference equations (3.1.23).

**b.** Repeat part **a** using the one-sided approximations (3.1.28) for $v'$.

**c.** Repeat parts **a** and **b** for the boundary conditions $v'(0) = 1$, $v(1) = 0$, and then $v'(0) + v(0) = 1$, $v'(1) + \frac{1}{2}v(1) = 0$.

**3.1.5.** Assume that the function $a$ of (3.1.31) is twice differentiable. By the change of variable $w(x) = \sqrt{a(x)}v(x)$, show that (3.1.31) can be replaced by an equation of the form $w''(x) = c(x)w(x) + f(x)$. Give an expression for $c(x)$ and show how it can be approximated numerically.

**3.1.6.** Suppose the equation (3.1.4) is defined on the interval $[\delta, \gamma]$. Show that the equations (3.1.8) are still correct provided that $h = (\gamma - \delta/(n+1))$ and $v_i$ is interpreted as the approximate solution at $x_i = \delta + ih$.

**3.1.7.** Assuming that $v$ is sufficiently differentiable, show that (3.1.35) is, in general, only a first-order approximation to $v''(0)$. However, if $\alpha = 0$, argue that it is second-order by assuming that $v$ has been extended outside the interval $[0, 1]$ so that $v(-h) = v(h)$.

**3.1.8.** Let $A$ and $B$ be the tridiagonal matrices

$$A = \begin{bmatrix} a_1 & b_1 & & & \\ c_1 & \ddots & \ddots & & \\ & \ddots & & & b_{n-1} \\ & & & c_{n-1} & a_n \end{bmatrix}, \qquad B = \begin{bmatrix} a_1 & \gamma_1 & & & \\ \gamma_1 & \ddots & \ddots & & \\ & \ddots & & & \gamma_{n-1} \\ & & & \gamma_{n-1} & a_n \end{bmatrix},$$

where $b_j c_j > 0$ and $\gamma_i = \sqrt{b_i c_i}$, $i = 1, \ldots, n-1$. Show that if

$$D = \operatorname{diag}\left(1, \frac{b_1}{c_1}, \frac{b_1 b_2}{c_1 c_2}, \ldots, \frac{b_1 \cdots b_{n-1}}{c_1 \cdots c_{n-1}}\right)$$

then $B = D^{1/2} A D^{-1/2}$, where $D^{1/2} = \operatorname{diag}(d_1^{1/2}, d_2^{1/2}, \ldots, d_n^{1/2})$.

**3.1.9.** Derive the approximations (3.1.5) and (3.1.21) by means of interpolation polynomials: let $l$ be the polynomial of degree 1 that satisfies $l(x \pm h) = v(x \pm h)$, and show that $l'(x)$ is the approximation (3.1.21). Then let $q$ be the quadratic polynomial that satisfies $q(x) = v(x)$ and $q(x \pm h) = v(x \pm h)$, and show that $q''(x)$ gives the approximation (3.1.5).

**3.1.10.** Assume that the $b_i$ in (3.1.22) are all equal to a constant $b$. Give a condition that ensures that the matrix of (3.1.23) can be symmetrized as in Exercise 3.1.8. Also, if the $c_i = 0$ in (3.1.22), derive an expansion for the quantities $d_n$ of Exercise 3.1.8 as $n \to \infty$.

**3.1.11.** Show that the matrix (3.1.46) can be written as $A = E^T D E$, where $E$ is the $(n+1) \times n$ matrix

$$E = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & & \ddots & \\ & & \ddots & 1 \\ & & & -1 \end{bmatrix}.$$

**3.1.12.** Let $A = I - B$, where $\|B\|_\infty < 1$ (See Appendix 2). Modify the proof of Theorem 3.1.1 to show that $A$ is nonsingular.

## 3.2 Solution of the Discretized Problem

In the previous section we saw that the use of finite difference discretization of the two-point boundary value problem (3.1.3) led to a system of linear equations. The exact form of this system depends on the boundary conditions,

but in all the cases we considered, except periodic boundary conditions, the system was of the tridiagonal form

$$\begin{bmatrix} a_{11} & a_{12} & & & \\ a_{21} & a_{22} & a_{23} & & \\ & a_{32} & \ddots & \ddots & \\ & & \ddots & & a_{n-1,n} \\ & & & a_{n,n-1} & a_{nn} \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ \\ \\ v_n \end{bmatrix} = \begin{bmatrix} d_1 \\ \vdots \\ \\ \\ d_n \end{bmatrix}. \tag{3.2.1}$$

### Gaussian Elimination

We will solve the system (3.2.1) by the Gaussian elimination method. This method, along with several variants, will be considered in detail for general linear systems in the next chapter. For the tridiagonal system (3.2.1) it turns out to be particularly simple. We consider first an example for $n = 3$ in which the equations are

$$\begin{aligned} 2v_1 - v_2 &= 1 \\ -v_1 + 2v_2 - v_3 &= 2 \\ -v_2 + 2v_3 &= 3\,. \end{aligned} \tag{3.2.2}$$

Multiply the first equation by $\frac{1}{2}$ and add to the second equation. This eliminates the coefficient of $v_1$ in the second equation, which becomes

$$\frac{3}{2}v_2 - v_3 = \frac{5}{2}. \tag{3.2.3}$$

Now multiply this equation by $\frac{2}{3}$ and add to the original third equation. This eliminates the coefficient of $v_2$ in the third equation and gives the modified equation

$$\frac{4}{3}v_3 = \frac{14}{3},$$

in which $v_3$ is the only unknown. Thus $v_3 = \frac{7}{2}$. Then (3.2.3) gives $v_2 = 4$, and the first equation of (3.2.2) gives $v_1 = \frac{5}{2}$.

The above example illustrates the main ideas of Gaussian elimination for the general tridiagonal system (3.2.1), which we write out in equation form as

$$\begin{aligned} a_{11}v_1 + a_{12}v_2 &= d_1 \\ a_{21}v_1 + a_{22}v_2 + a_{23}v_3 &= d_2 \\ a_{32}v_2 + a_{33}v_3 + a_{34}v_4 &= d_3 \\ \ddots & \vdots \\ a_{nn-1}v_{n-1} + a_{nn}v_n &= d_n. \end{aligned} \tag{3.2.4}$$

We assume that $a_{11} \neq 0$, multiply the first equation by $a_{21}/a_{11}$, and subtract it from the second. This eliminates the coefficient of $v_1$ in the second equation so that the last $n-1$ equations are

$$\begin{aligned} a'_{22}v_2 + a_{23}v_3 & & &= d'_2 \\ a_{32}v_2 + a_{33}v_3 & + a_{34}v_4 & &= d_3 \\ & \ddots & & \vdots \\ & & a_{nn-1}v_{n-1} + a_{nn}v_n &= d_n, \end{aligned}$$

where the primes indicate that the elements have been changed by the first row operations. This system is again tridiagonal, and we repeat the process: multiply the first row by $a_{32}/a'_{22}$ and subtract from the next row to eliminate the coefficient of $v_2$. Note that in this step, as in all subsequent ones, we must assume that the divisor element is non-zero. We will consider this assumption in much more detail in the next chapter.

At the end of the above process, the system has been transformed to

$$\begin{aligned} a_{11}v_1 + a_{12}v_2 & & &= d_1 \\ a'_{22}v_2 + a_{23}v_3 & & &= d'_2 \\ \ddots & & \vdots & \\ & a'_{n-1,n-1}v_{n-1} + a_{n-1,n}v_n &= d'_{n-1} \\ & a'_{n,n}v_n &= d'_n. \end{aligned} \tag{3.2.5}$$

Since the last equation contains only a single unknown, it can be solved. Then with $v_n$ computed, the next to last equation can be solved for $v_{n-1}$, and so on, working our way back up the equations. This process is known as the *back substitution*, at the end of which we have computed the solution of the system (3.2.4).

As we shall see in the next chapter, Gaussian elimination for a general linear system of $n$ equations requires approximately $\frac{2}{3}n^3$ arithmetic operations. Because of the simple form of tridiagonal systems, far fewer operations are needed. To reduce the original system (3.2.4) to (3.2.5) requires only $2(n-1)$ additions, $2(n-1)$ multiplications, and $n-1$ divisions, while the back substitution requires $n-1$ additions, $n-1$ multiplications, and $n$ divisions. Hence the total operation count to solve a tridiagonal system is

$$3(n-1) \text{ additions } + \ 3(n-1) \text{ multiplications } + \ (2n-1) \text{ divisions.}$$

Because of this relatively small operation count, very large systems can be solved very rapidly. Suppose, for example, that multiplication and addition each require $1\mu s$ ($=10^{-6}$ seconds) and division requires $5\mu s$. Then the arithmetic operation time for $n = 10,000$ is approximately

$$(60,000 + 20,000 \times 5)\mu s = 0.16 \text{ seconds.}$$

### A Numerical Example

To illustrate the methods of this chapter we solve a particular two-point boundary value problem numerically. The equation is

$$[(1+x^2)v']' = 2 + 6x^2 + 2x\cos x - (1+x^2)\sin x, \tag{3.2.6}$$

with the boundary conditions

$$v(0) = 1, \qquad v(1) = 2 + \sin 1. \tag{3.2.7}$$

To evaluate the accuracy of the methods we have chosen the right hand side of (3.2.6) so that the exact solution is known:

$$v(x) = x^2 + \sin x + 1. \tag{3.2.8}$$

The equation (3.2.6) is of the form (3.1.31), but we can carry out the differentiation to write it as

$$(1+x^2)v'' + 2xv' = d(x), \tag{3.2.9}$$

where

$$d(x) = 2 + 6x^2 + 2x\cos x - (1+x^2)\sin x. \tag{3.2.10}$$

We could divide (3.2.9) through by $1+x^2$ to obtain the form (3.1.3). Alternatively, we can modify the difference equations by multiplying the approximation of $v''$ by $1+x^2$; that is,

$$(1+x^2)v''\,\big|_{x_i} \doteq \frac{(1+x_i^2)}{h^2}(v_{i+1} - 2v_i + v_{i-1}). \tag{3.2.11}$$

Using (3.2.11) and central differences to approximate $v'$, the difference equations have the matrix form (3.1.23), where if $a_i = 1 + x_i^2$

$$p_i = 2a_i, \quad q_i = -a_i - x_i h, \quad r_i = -a_i + x_i h. \tag{3.2.12}$$

Since $x_i = ih$, and $a_i = 1 + (ih)^2$, the coefficient matrix is then

$$\begin{bmatrix} 2+2h^2 & -1-2h^2 & & & \\ -1-2h^2 & 2+8h^2 & -1-6h^2 & & \\ \ddots & \ddots & \ddots & & \\ & & & -1-n(n-1)h^2 & \\ & & -1-n(n-1)h^2 & 2+2n^2h^2 & \end{bmatrix}, \tag{3.2.13}$$

and the right hand side is

$$-h^2\left\{d_1 - \frac{1}{h^2}, d_2, \ldots, d_{n-1}, d_n - \left[n(n+1) + \frac{1}{h^2}\right](2 + \sin 1)\right\}^T. \tag{3.2.14}$$

On the other hand, if we use the one-sided differences (3.1.28) to approximate $v'$, the coefficient matrix and the right hand side are

$$\begin{bmatrix} 2+4h^2 & -1-3h^2 & & & \\ -1-4h^2 & 2+12h^2 & -1-8h^2 & & \\ \ddots & \ddots & \ddots & & \\ & & & -1-(n-1)^2h^2 & \\ & & -1-n^2h^2 & 2+2n(n+1)h^2 & \end{bmatrix} \tag{3.2.15}$$

$$-h^2\left[d_1-\left(1+\frac{1}{h^2}\right),d_2,\ldots,d_n-\left(n^2+2n+\frac{1}{h^2}\right)(2+\sin 1)\right]^T. \tag{3.2.16}$$

It is left to Exercise 3.2.2 to verify (3.2.13) - (3.2.16).

We could also use the discretization of (3.2.6) discussed in the Supplementary Discussion of Section 3.1. In this case the matrix (3.1.46) becomes

$$\begin{bmatrix} 2+\frac{5}{2}h^2 & -1-\frac{9}{4}h^2 & & & \\ -1-\frac{9}{4}h^2 & 2+\frac{17}{2}h^2 & -1-\frac{25}{4}h^2 & & \\ \ddots & \ddots & \ddots & & \\ & & & -1-(n-\frac{1}{2})^2h^2 & \\ & & -1-(n-\frac{1}{2})^2h^2 & 2+(2n^2+\frac{1}{2})h^2 & \end{bmatrix}, \tag{3.2.17}$$

and the right-hand side is

$$-h^2\left\{d_1-\left(\frac{1}{4}+\frac{1}{h^2}\right),d_2,\ldots,d_n-\left[(n+\tfrac{1}{2})^2+\frac{1}{h^2}\right](2+\sin 1)\right\}^T. \tag{3.2.18}$$

All three systems were solved for various values of $h$. Figure 3.2 shows the approximate solution using (3.2.13) and (3.2.14) with a step size $h = 1/32$. This agrees very closely with the exact solution (3.2.8). The other methods generated visibly indistinguishable approximate solutions.

### Periodic Boundary Conditions

We conclude this chapter by considering other boundary conditions. As discussed in Section 3.1, Neumann boundary conditions also lead to tridiagonal systems of equations, but periodic boundary conditions introduce matrix elements outside of the tridiagonal part, as shown in (3.1.41). We could solve linear systems with such coefficient matrices by Gaussian elimination, but we wish to consider an alternative.

For any non-zero column vectors $\mathbf{u}$ and $\mathbf{v}$ of length $n$, the product $\mathbf{u}\mathbf{v}^T$ is called a *rank-one* matrix and is an $n \times n$ matrix whose $i,j$ entry is $u_i v_j$. Now

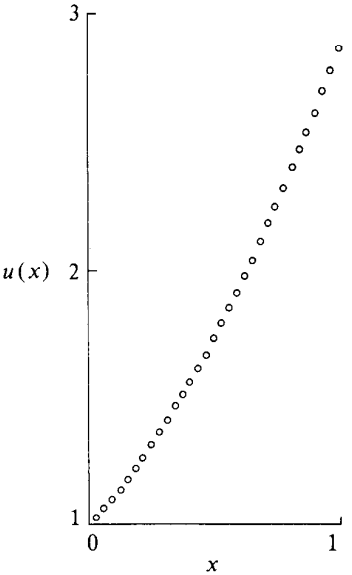


Figure 3.2: *The Computed Solution to* (3.2.9), (3.2.7), *using* $h = 1/32$ *for the Difference Method Corresponding to* (3.2.13), (3.2.14)

let $A$ be the matrix of (3.1.41) and $B$ the tridiagonal part of the matrix. Then $A$ may be written as (Exercise 3.2.7)

$$A = B - \mathbf{e}_1\mathbf{e}_n^T - \mathbf{e}_n\mathbf{e}_1^T, \tag{3.2.19}$$

where $\mathbf{e}_i$ is the vector with 1 in the $i$th position and zero elsewhere. Thus $A$ may be written as the sum of its tridiagonal part plus two rank-one matrices that bring in the outlying $-1$'s. We may also write $A$ as

$$A = T - \mathbf{w}\mathbf{w}^T, \tag{3.2.20}$$

where $T = B + \text{diag}(1, 0, \cdots, 0, 1)$ and $\mathbf{w} = \mathbf{e}_1 + \mathbf{e}_n$. This changes the original tridiagonal matrix but has the advantage that $A$ is now a tridiagonal matrix plus a single rank-one matrix.

Next let $C$ be any nonsingular matrix and $\mathbf{u}\mathbf{v}^T$ a rank-one matrix. Then the *Sherman-Morrison* formula is

$$(C + \mathbf{u}\mathbf{v}^T)^{-1} = C^{-1} - \alpha^{-1}C^{-1}\mathbf{u}\mathbf{v}^TC^{-1}, \quad \alpha = 1 + \mathbf{v}^TC^{-1}\mathbf{u}, \tag{3.2.21}$$

which is easily verified (Exercise 3.2.7). The condition for nonsingularity of $C + \mathbf{u}\mathbf{v}^T$ is that $\alpha \neq 0$. Note that the matrix added to $C^{-1}$ on the right side of (3.2.21) is also a rank-one matrix. To solve a linear system of the form

$$(C + \mathbf{u}\mathbf{v}^T)\mathbf{x} = \mathbf{b}, \tag{3.2.22}$$

we do not wish to form the inverse as in (3.2.21), but only solve linear systems. Thus we use (3.2.21) to write the solution of (3.2.22) in the form

$$\begin{aligned} \mathbf{x} &= (C + \mathbf{u}\mathbf{v}^T)^{-1}\mathbf{b} = C^{-1}\mathbf{b} - \alpha^{-1}C^{-1}\mathbf{u}\mathbf{v}^T C^{-1}\mathbf{b} \qquad (3.2.23) \\ &= \mathbf{y} - \alpha^{-1}(\mathbf{v}^T\mathbf{y})\mathbf{z}, \qquad \alpha = 1 + \mathbf{v}^T\mathbf{z}, \end{aligned}$$

where $\mathbf{y}$ is the solution of $C\mathbf{y} = \mathbf{b}$ and $\mathbf{z}$ the solution of $C\mathbf{z} = \mathbf{u}$. In particular, if $C$ is a tridiagonal matrix, we can solve the system (3.2.22) by solving two tridiagonal systems and then combining those solutions as shown in (3.2.23). Thus, for the matrix of (3.2.20), the solution of $A\mathbf{x} = \mathbf{b}$ would consist of the steps:

$$\text{Solve } T\mathbf{y} = \mathbf{b},\ T\mathbf{z} = \mathbf{w}, \qquad (3.2.24)$$

$$\text{Form } \alpha = 1 + \mathbf{w}^T\mathbf{z},\ \mathbf{x} = \mathbf{y} - \alpha^{-1}(\mathbf{w}^T\mathbf{y})\mathbf{z}. \qquad (3.2.25)$$

In the next chapter we will see (Exercises 4.2.20,21) that this approach has a slightly lower operation count than applying Gaussian elimination to $A$ itself. More importantly, it requires only a code for solving tridiagonal systems (plus the additional operations of (3.2.25)) whereas a Gaussian elimination code that takes advantage of the zeros in $A$ is somewhat more complicated.

Although we were able to convert the matrix $A$ of (3.2.19) to the form (3.2.20), which involved only a single rank-one matrix, in many situations we wish to deal with a matrix of the form $C + R$, where $R$ is a matrix of rank $m$. A rank $m$ matrix may be written in the form $R = UV^T$, where $U$ and $V$ are $n \times m$ matrices. Then (3.2.21) extends to the *Sherman-Morrison-Woodbury* formula

$$(C + UV^T)^{-1} = C^{-1} - C^{-1}U(I + V^T C^{-1}U)^{-1}V^T C^{-1}. \qquad (3.2.26)$$

The matrix $I + V^T C^{-1}U$ is $m \times m$, and the Sherman-Morrison formula (3.2.21) is the special case $m = 1$ of (3.2.26). We could apply (3.2.26) to solve the system $A\mathbf{x} = \mathbf{b}$, where $A$ is given by (3.2.19), although it is slightly more efficient to use (3.2.20) and (3.2.21). To use the formula (3.2.26) we take $C = B$, $U = (\mathbf{e}_1, \mathbf{e}_n)$;and $V = -(\mathbf{e}_n, \mathbf{e}_1)$. The details of the computation are left to Exercise 3.2.8.

## Supplementary Discussion and References: 3.2

For a nice review of the history and many applications of the formulas (3.2.21) and (3.2.26), see Hager [1989]. In particular, (3.2.21) was first given by J. Sherman and W. Morrison in 1949 for the special case of changing the elements in one column of $C$; in this case $\mathbf{v} = \mathbf{e}_i$ if the $i$th column is changed. The general formula (3.2.21) was given by M. Bartlett in 1951. Simultaneously,

M. Woodbury gave the still more general formula (3.2.26) in a 1950 report, although it had appeared in earlier work in the mid 1940's. For a discussion of rounding error analysis of these formulas, see Yip [1986].

## EXERCISES 3.2

**3.2.1.** Solve the linear system

$$\begin{bmatrix} 4 & 3 & 0 \\ 3 & 5 & 4 \\ 0 & 4 & 6 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

by Gaussian elimination.

**3.2.2.** Verify that the coefficient matrix and right hand side of the difference approximation of (3.2.9) using (3.2.11) and central difference approximations for $v'$ are given by (3.2.13) and (3.2.14). Verify (3.2.15) and (3.2.16) if one-sided approximations to $v'$ are used.

**3.2.3.** Write a computer program to implement Gaussian elimination for tridiagonal systems. Use one-dimensional arrays to store the matrix.

**3.2.4.** Multiply two $n \times n$ tridiagonal matrices. How many arithmetic operations does this require? Is the product matrix tridiagonal?

**3.2.5.** Solve numerically the system with coefficient matrix and right-hand side (3.2.13), (3.2.14) for $n = 10$ and $n = 20$. Discuss the accuracy of your approximate solutions.

**3.2.6.** Repeat Exercise 3.2.5 for (3.2.15), (3.2.16) and (3.2.17), (3.2.18).

**3.2.7.** Verify formulas (3.2.19), (3.2.21), and (3.2.26).

**3.2.8.** Verify that the formula (3.2.26) can be used to solve the system $A\mathbf{x} = \mathbf{b}$, where $A$ is given by (3.2.19), by carrying out the following steps:

1. Solve $B\mathbf{y} = \mathbf{b}$, $B\mathbf{w}_1 = \mathbf{e}_1$, $B\mathbf{w}_n = \mathbf{e}_n$.
2. Form the $2 \times 2$ matrix

$$U^T W = \begin{pmatrix} \mathbf{e}_1^T \\ \mathbf{e}_n^T \end{pmatrix} (\mathbf{w}_1, \mathbf{w}_n) = \begin{pmatrix} \mathbf{e}_1^T \mathbf{w}_1 & \mathbf{e}_1^T \mathbf{w}_n \\ \mathbf{e}_n^T \mathbf{w}_1 & \mathbf{e}_n^T \mathbf{w}_n \end{pmatrix}.$$

3. Form the 2-vector

$$\mathbf{q} = (I + U^T W)^{-1} U^T \mathbf{y}.$$

4. Form the solution

$$\mathbf{x} = \mathbf{y} - W\mathbf{q}.$$

**3.2.9.** Consider the problem on the infinite interval $(0, \infty)$:

$$y''(x) = (2x + 2)/(x + 2)y(x),\ y(0) = 2,\ y(\infty) = 0.$$

**a.** Approximate a solution to this problem by replacing the boundary condition at $\infty$ by $y(5) = 0$ and using finite differences for $h = 1/10, 1/20$ and $1/32$.

**b.** Now note that $(2x+2)/(x+2) \to 2$ as $x \to \infty$. Hence consider the problem

$$z''(x) = 2z(x),\ z(0) = 2,\ z(\infty) = 0.$$

Solve this problem exactly and find an expression for $z'(x)$ in terms of $z(x)$. Using this expression for the value of $y'(x)$ as $x \to \infty$, solve the truncated tridiagonal system and compare the solution to the truncated system used in part a.