1
Background Mathematics

1.1 Introduction

The companion volume of this book, 'Analytic methods for partial differential equations', is concerned with solution of partial differential equations using classical methods which result in analytic solutions. These equations result when almost any physical situation is modelled, ranging from fluid mechanics problems, through electromagnetic problems to models of the economy. Some specific and fundamental problems were highlighted in the earlier volume, namely the three major classes of linear second order partial differential equations. The heat equation, the wave equation and Laplace's equation will form a basis for study from a numerical point of view for the same reason as they did in the analytic case. That is, the three equations are the canonical forms to which any quasi-linear second order equation may be reduced using the characteristic transformation.

The history of the numerical solution of partial differential equations is much more recent than the analytic approaches, and the development of the numerical approach has been heavily influenced by the advent of high speed computing machines. This progress is still being seen. In the pre-computer days, the pressures of war were instrumental in forcing hand-worked numerical solutions to problems such as blast waves to be attempted. Large numbers of electro-mechanical machines were used with the 'programmer' controlling the machine operators. Great ingenuity was required to allow checks to be made on human error in the process. The relaxation method was one of the results of these processes.

Once reliable and fast electronic means were available, the solution of more

and more complex partial differential equations became feasible. The earliest method involved discretising the partial derivatives and hence converting the partial differential equation into a difference equation. This could either be solved in a step-by-step method as with the wave equation, or required the solution of a large set of sparse linear algebraic equations in the case of Laplace's equation. Hence speed was not the only requirement of the computing machine. Storage was also crucial. At first, matrix blocks were moved in and out of backing store to be processed in the limited high speed store available. Today, huge storage requirements can be met relatively cheaply, and the progress in cheap high speed store and fast processing capability is enabling more and more difficult problems to be attempted. Weather forecasting is a very well-known area in which computer power is improving the accuracy of forecasts: admittedly now combined with the knowledge of chaos which gives some degree of forecast reliability.

In the chapters which follow the numerical solution of partial differential equations is considered, first by using the three basic problems as cases which demonstrate the methods. The finite difference method is considered first. This is the method which was first applied in the early hand-computed work, and is relatively simple to set up. The area or volume of interest is broken up into a grid system on which the partial derivatives can be expressed as simple differences. The problem then reduces to finding the solution at the grid points as a set of linear algebraic equations. Hence some attention is paid to solving linear algebraic equations with many elements in each row being zero. The use of iterative solutions can be very effective under these circumstances.

A number of theoretical considerations need to be made. Firstly, it needs to be established that by taking a finer and finer grid, the difference equation solution does indeed converge to the solution of the approximated partial differential equation. This is the classical problem of accuracy in numerical analysis. However real computers execute their arithmetic operations to a finite word length and hence all stored real numbers are subject to a rounding error. The propagation of these errors is the second main theme of numerical analysis in general, and partial differential equations in particular. This is the problem of numerical stability. Do small errors introduced as an inevitable consequence of the use of finite word length machines grow in the development of the solution? Questions of this sort will be considered as part of the stability analysis for the methods presented. There will be exercises in which the reader will be encouraged to see just how instability manifests itself in an unstable method, and how the problem can be circumvented.

Using finite differences is not the only way to tackle a partial differential equation. In 1960, Zienkiewicz used a rather different approach to structural problems in civil engineering, and this work has developed into a completely separate method of solution. This method is the finite element method (Zienkiewicz, 1977). It is based on a variational formulation of the partial differential equation, and the first part of the description of the method

requires some general ways of obtaining a suitable variational principle. In many problems there is a natural such principle, often some form of energy conservation. The problem is then one of minimising an integral by the choice of a dependent function. The classic method which then follows is the Rayleigh–Ritz method. In the finite element method, the volume over which the integral is taken is split up into a set of elements. These may be triangular or prismatic for example. On each element a simple solution form may be assumed, such as a linear form. By summing over each element, the condition for a minimum reduces to a large set of linear algebraic equations for the solution values at key points of the element, such as the vertices of the triangle. Again the use of sparse linear equation solvers is required.

This first chapter is concerned with some of the mathematical preliminaries which are required in the numerical work. For the most part this chapter is quite independent of the equivalent chapter in the first volume, but the section on classification reappears here for completeness.

1.2 Vector and Matrix Norms

The numerical analysis of partial differential equations requires the use of vectors and matrices both in setting up the numerical methods and in analysing their convergence and stability properties. There is a practical need for measures of the 'size' of a vector or matrix which can be realised computationally, as well as be used theoretically. Hence the first section of this background chapter deals with the definition of the norm of a vector, the norm of a matrix and realises some specific examples.

The norm of vector $\mathbf{x}$ is a real positive number, $||\mathbf{x}||$, which satisfies the axioms:

(i) $||\mathbf{x}|| > 0$ if $\mathbf{x} \neq \mathbf{0}$ and $||\mathbf{x}|| = 0$ if $\mathbf{x} = \mathbf{0}$;

(ii) $||c\mathbf{x}|| = |c|\,||\mathbf{x}||$ for a real or complex scalar c; and

(iii) $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$.

If the vector $\mathbf{x}$ has components $x_1, \ldots, x_n$ then there are three commonly used norms:

(i) The one-norm of $\mathbf{x}$ is defined as

$$||\mathbf{x}||_1 = |x_1| + |x_2| + \cdots + |x_n| = \sum_{i=1}^{n} |x_i|. \tag{1.2.1}$$

(ii) The two-norm of $\mathbf{x}$ is

$$||\mathbf{x}||_2 = (|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2)^{\frac{1}{2}} = \left[\sum_{i=1}^{n} |x_i|^2\right]^{\frac{1}{2}}. \tag{1.2.2}$$

(iii) The infinity norm of $\mathbf{x}$ is the maximum of the moduli of the components or

$$||\mathbf{x}||_\infty = \max_i |x_i|. \tag{1.2.3}$$

In a similar manner the norm of a matrix A is a real positive number giving a measure of the 'size' of the matrix which satisfies the axioms

(i) $||A|| > 0$ if $A \neq 0$ and $||A|| = 0$ if $A = 0$;

(ii) $||cA|| = |c|\,||A||$ for a real or complex scalar c;

(iii) $||A + B|| \leq ||A|| + ||B||$; and

(iv) $||AB|| \leq ||A||\,||B||$.

Vectors and matrices occur together and so they must satisfy a condition equivalent to (iv), and with this in mind matrix and vector norms are said to be *compatible* or *consistent* if

$$||A\mathbf{x}|| \leq ||A||\,||\mathbf{x}||, \quad \forall x \in \mathcal{R}^n (\text{or } \mathcal{C}^n).$$

There is a class of matrix norms whose definition depends on an underlying vector norm. These are the subordinate matrix norms. Let A be an $n \times n$ matrix and $\mathbf{x} \in S$ where

$$S = \{(n \times 1) \text{ vectors } : ||\mathbf{x}|| = 1\};$$

now in general $||A\mathbf{x}||$ varies as $\mathbf{x}$ varies ($\mathbf{x} \in S$). Let $\mathbf{x}_0 \in S$ be such that $||A\mathbf{x}||$ attains its maximum value. Then the norm of matrix A, subordinate to the vector norm $||.||$, is defined by

$$||A|| = ||A\mathbf{x}_0|| = \max_{||\mathbf{x}||=1} ||A\mathbf{x}||. \tag{1.2.4}$$

The matrix norm that is subordinate to the vector norm automatically satisfies the compatibility condition since, if $\mathbf{x} = \mathbf{x}_1 \in S$, then

$$||A\mathbf{x}_1|| \leq ||A\mathbf{x}_0|| = ||A|| = ||A||\,||\mathbf{x}_1|| \quad \text{since} \quad ||\mathbf{x}_1|| = 1.$$

Therefore $||A\mathbf{x}|| \leq ||A||\,||\mathbf{x}||$ for any $\mathbf{x} \in \mathcal{R}^n$. Note that for all subordinate matrix norms

$$||I|| = \max_{||\mathbf{x}||=1} ||I\mathbf{x}|| = \max_{||\mathbf{x}||=1} ||\mathbf{x}|| = 1. \tag{1.2.5}$$

The definitions of the subordinate one, two and infinity norms with $||\mathbf{x}|| = 1$ lead to:

- The one norm of matrix A is the maximum column sum of the moduli of the elements of A, and is denoted by $||A||_1$.
- The infinity norm of matrix A is the maximum row sum of the moduli of the elements of A, and is denoted by $||A||_\infty$.
- The two norm of matrix A is the square root of the spectral radius of $A^H A$ where $A^H = (\bar{A})^T$ (the transpose of the complex conjugate of A). This norm is denoted by $||A||_2$. The *spectral radius* of a matrix B is denoted by $\rho(B)$ and is the modulus of the eigenvalue of maximum modulus of B.

Hence for example if

$$A = \begin{pmatrix} -1 & 1 \\ 3 & -2 \end{pmatrix} \quad \text{then} \quad A^H A = A^T A = \begin{pmatrix} 10 & -7 \\ -7 & 5 \end{pmatrix}$$

has eigenvalues 14.93 and 0.067. Then using the above definitions

$$||A||_1 = 1 + 3 = 4, \; ||A||_\infty = 3 + 2 = 5, \; ||A||_2 = \sqrt{14.93} = 3.86.$$

Note that if A is real and symmetric

$$A^H = A \quad \text{and} \quad ||A||_2 = [\rho(A^2)]^{\frac{1}{2}} = [\rho^2(A)]^{\frac{1}{2}} = \rho(A) = \max_i |\lambda_i|.$$

A number of other equivalent definitions of $||A||_2$ appear in the literature. For example the eigenvalues of $A^H A$ are denoted by $\sigma_1^2, \sigma_2^2, \ldots \sigma_n^2$ and the σ_i are called the *singular values* of A. By their construction the singular values will be real and non-negative. Hence from the above definition

$$||A||_2 = \sigma_1$$

where

$$\sigma_1 = \max_{1 \le i \le n} \sigma_i.$$

For a symmetric A, the singular values of A are precisely the eigenvalues of A apart from a possible sign change, and

$$||A||_2 = |\lambda_1|,$$

where λ_1 is the largest absolute value eigenvalue of A. A bound for the spectral radius can also be derived in terms of norms. Let λ_i and $\mathbf{x}_i$ be corresponding eigenvalue and eigenvector of the $n \times n$ matrix A, then $A\mathbf{x}_i = \lambda_i \mathbf{x}_i$, and

$$||A\mathbf{x}_i|| = ||\lambda_i \mathbf{x}_i|| = |\lambda_i| \, ||\mathbf{x}_i||.$$

For all compatible matrix and vector norms

$$|\lambda_i| \, ||\mathbf{x}_i|| = ||A\mathbf{x}_i|| \le ||A|| \, ||\mathbf{x}_i||.$$

Therefore $|\lambda_i| \le ||A||$, $i = 1(1)n$. Hence $\rho(A) \le ||A||$ for any matrix norm that is compatible with a vector norm.

A few illustrative exercises which are based on the previous section now follow.

EXERCISES

1.1 A further matrix norm is $||A||_E$ or the Euclidean norm, defined by

$$||A||_E^2 = \sum_{i,j} a_{ij}^2.$$

Prove that

$$||A||_2 \leq ||A||_E \leq n^{1/2}||A||_2$$

for an nth order matrix A. Verify the inequality for the matrix

$$\begin{bmatrix} 1.2 & -2.0 \\ 1.0 & 0.6 \end{bmatrix}.$$

1.2 Compute the four norms $||A||_1$, $||A||_2$, $||A||_\infty$ and $||A||_E$ for the matrix

$$\begin{bmatrix} 1 & 0 & 1 \\ 2 & 3 & 0 \\ 2 & 1 & 4 \end{bmatrix}$$

and find the characteristic polynomial of this matrix and hence its eigenvalues. Verify that

$$|\lambda_i| \leq ||A||$$

for $i = 1, 2, 3$.

1.3 Compute the spectral radius of the matrix

$$\begin{bmatrix} 9 & -2 & 1 \\ 4 & 5 & -2 \\ 1 & -3 & -5 \end{bmatrix}$$

and confirm that this value is indeed less than or equal to both $||A||_1$ and $||A||_\infty$.

1.4 For the solution of a set of linear algebraic equations

$$A\mathbf{x} = \mathbf{b}$$

the condition number is given by

$$\kappa = ||A||_2||A^{-1}||_2.$$

The input errors, such as machine precision, are multiplied by this number to obtain the errors in the solution x. Find the condition number for the matrix

$$\begin{bmatrix} 1 & 2 & 6 \\ 2 & 6 & 24 \\ 6 & 24 & 120 \end{bmatrix}.$$

1.5 Sketch the curve of $\kappa(A)$ as defined above, against the variable c for the matrix

$$\begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & c \end{bmatrix}.$$

Large values of κ will indicate where there will be high loss of accuracy in the solution of linear equations with the matrix A. For $c = 1/5$ the matrix is the 3×3 Hilbert matrix.

1.3 Gerschgorin's Theorems

The first of the theorems which gives bounds on eigenvalues is

Theorem 1.1

The largest of the moduli of the eigenvalues of the square matrix A cannot exceed the largest sum of the moduli of the elements along any row or any column. In other words $\rho(A) \leq ||A||_1$, or $||A||_\infty$.

Proof

Let λ_i be an eigenvalue of the $n \times n$ matrix A and $\mathbf{x}_i$ be the corresponding eigenvector with components $v_1, v_2, \ldots, v_n$. Then $A\mathbf{x}_i = \lambda_i \mathbf{x}_i$ becomes in full

$$\begin{aligned} a_{11}v_1 + a_{12}v_2 + \cdots + a_{1n}v_n &= \lambda_i v_1 \\ a_{21}v_1 + a_{22}v_2 + \cdots + a_{2n}v_n &= \lambda_i v_2 \\ \vdots &= \vdots \\ a_{s1}v_1 + a_{s2}v_2 + \cdots + a_{sn}v_n &= \lambda_i v_s \\ \vdots &= \vdots \end{aligned}$$

Let v_s be the largest in modulus of $v_1, \ldots, v_n$, noting that $v_s \neq 0$. Select the sth equation and divide by v_s giving

$$\lambda_i = a_{s1}\left(\frac{v_1}{v_s}\right) + a_{s2}\left(\frac{v_2}{v_s}\right) + \cdots + a_{sn}\left(\frac{v_n}{v_s}\right).$$

Therefore

$$|\lambda_i| \leq |a_{s1}| + |a_{s2}| + \cdots + |a_{sn}|$$

since

$$\left|\frac{v_i}{v_s}\right| \leq 1, \quad i = 1, 2, \ldots, n.$$

If this is not the largest row sum then $|\lambda_i| <$ the largest row sum. In particular this holds for

$$|\lambda_i| = \max_{1 \le s \le n} |\lambda_s|.$$

Since the eigenvalues of A^T are the same as those for A the theorem is also true for columns.

The second theorem gives the approximate position of the eigenvalues of a matrix and is *Gerschgorin's circle theorem or first theorem, or Brauer's theorem.*

Theorem 1.2

Let P_s be the sum of the moduli of the elements along the sth row excluding the diagonal element a_{ss}. Then each eigenvalue of A lies inside or on the boundary of at least one of the circles $|\lambda - a_{ss}| = P_s$.

Proof

By the previous proof

$$\lambda_i = a_{s1}\left(\frac{v_1}{v_s}\right) + a_{s2}\left(\frac{v_2}{v_s}\right) + \cdots + a_{ss} + \cdots + a_{sn}\left(\frac{v_n}{v_s}\right).$$

Hence

$$|\lambda_i - a_{ss}| = \sum_{j \neq s} a_{sj}\frac{v_j}{v_s} = \sum_{j \neq s} |a_{sj}|.$$

The third theorem is *Gerschgorin's second theorem*:

Theorem 1.3

If p of the circles of Gerschgorin's circle theorem form a connected domain that is isolated from the other circles, then there are precisely p eigenvalues of the matrix A within this connected domain. In particular, an isolated Gerschgorin's circle contains one eigenvalue.

Proof

Split the matrix A into its diagonal elements D and off-diagonal elements C to give

$$A = \text{diag}(a_{ii}) + C = D + C \tag{1.3.1}$$

and then using P_s as defined in the circle theorem, the matrices $(D+\epsilon C)$ can be considered. For $\epsilon = 0$ the matrix D is returned whereas $\epsilon = 1$ gives A. However the characteristic polynomial of $(D+\epsilon C)$ has coefficients which are themselves

polynomials in ϵ, and therefore the roots of this characteristic polynomial are continuous in ϵ. Hence eigenvalues traverse continuous paths as ϵ varies, by the circle theorem as, for any eigenvalue, the eigenvalues lie in circular discs with centres a_{ii} and radii ϵP_i.

Suppose the first s discs form a connected domain. The discs may be reordered if this is not the case. Then $(n-s)$ discs with radii $P_{s+1}, P_{s+2}, \ldots, P_n$ are isolated from the s with radii $P_1, P_2, \ldots, P_s$. This also applies to the discs of radii ϵP_i for all $\epsilon \in [0,1]$. When $\epsilon = 0$, the eigenvalues are $a_{11}, \ldots, a_{nn}$ and clearly the first s lie inside the domain corresponding to the first s discs, and the other $(n-s)$ lie outside. Hence by the continuity this state must continue through to $\epsilon = 1$ to prove the theorem.

When the eigenvalues λ_i of a matrix A are estimated by the circle theorem, the condition $|\lambda_i| \leq 1$ is equivalent to $||A||_\infty \leq 1$ or $||A||_1 \leq 1$, for we have $|\lambda - a_{ss}| \leq P_s$. Hence $-P_s \leq \lambda - a_{ss} \leq P_s$ so that $-P_s + a_{ss} \leq \lambda \leq P_s + a_{ss}$. Now λ will satisfy $-1 \leq \lambda \leq 1$, if $P_s - a_{ss} \leq 1$ and $P_s + a_{ss} \leq 1$ for $s = 1, \ldots, n$, as

$$||A||_\infty = \max_{1 \leq s \leq n} \sum_{j=1}^{n} |a_{sj}| = P_s + |a_{ss}| \leq 1. \tag{1.3.2}$$

Now P_s is the sum of the moduli of the elements of A in the sth row (excluding a_{ss}), and a_{ss} may be positive or negative. Hence inequality (1.3.2) is equivalent to

$$\sum_{j=1}^{n} |a_{sj}| \leq 1, \quad s = 1, \ldots, n, \tag{1.3.3}$$

or to $||A||_\infty \leq 1$ for rows. For any consistent pairs of matrix and vector norms,

$$|\lambda| \, ||\mathbf{x}|| = ||\lambda \mathbf{x}|| = ||A\mathbf{x}|| \leq ||A|| \, ||\mathbf{x}||$$

and hence $|\lambda| \leq ||A||$. Hence

$$||A||_2^2 = \text{max eigenvalue of} \quad A^H A$$

$$\leq ||A^H A||_1 \leq ||A^H||_1 ||A||_1 = ||A||_\infty ||A||_1.$$

Hence $||A||_2^2 \leq ||A||_1 ||A||_\infty$, and so if both inequalities hold it follows automatically that $||A||_2 \leq 1$. Hence Gerschgorin's circle theorem can be used to establish conditions for stability which will be dealt with in Chapter 2.

[Note: It can be shown that if the off-diagonal elements of a real tri-diagonal matrix are one-signed then all its eigenvalues are real (Smith, 1978).]

The following exercises can now be attempted to complete the understanding of this section.

EXERCISES

1.6 Use Gerschgorin's theorems to investigate the regions in which the eigenvalues of the matrix

$$\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

lie, and confirm that they lie in the range $[0, 4]$.

1.7 Use Gerschgorin's theorems to investigate whether an estimate of the condition number κ for the matrix of exercise 1.4 can be found. For a symmetric matrix, κ reduces to the ratio of the eigenvalue of largest modulus to that with the smallest modulus. This exercise highlights the shortcomings of the theorems.

1.8 Consider the same problem as in exercise 1.7. but with the $n \times n$ matrix whose elements are defined by

$$a_{i,j} = (i + j - 1)!.$$

This is an example of a well-known class of very ill-conditioned matrices.

1.9 Use Gerschgorin's theorems to find bounds on the condition factor $\kappa(A)$ for the matrix

$$\begin{bmatrix} 1 & 0.2 & 0.3 & 0 \\ 1 & 8 & 1 & 0 \\ 0 & 1 & 10 & 4 \\ 0 & 0 & 4 & 100 \end{bmatrix}.$$

1.4 Iterative Solution of Linear Algebraic Equations

In general the use of difference methods for the solution of partial differential equations leads to an algebraic system $A\mathbf{x} = \mathbf{b}$ where A is a given matrix which is sparse and of large order. Direct methods for solving this system, such as Gaussian elimination (Evans, 1995) tend to be inefficient and it is more usual to use an iterative method. The problem with elimination methods is that the initially sparse matrix begins to fill-in as the process develops, and more and more of the originally zero elements now have to be processed. There is a consequential cost in both storage and processor time. Iterative methods do

not alter the matrix structure and so preserve sparseness, though in some cases problems of convergence may become an issue.

Consider splitting the matrix A into three components in which L is lower triangular, D is diagonal and U is upper triangular. Let us suppose that

$$A = \begin{bmatrix} d_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ l_{21} & d_{22} & u_{23} & \dots & u_{2n} \\ l_{31} & l_{32} & d_{33} & \dots & u_{3n} \\ \vdots & & & & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & d_{nn} \end{bmatrix}.$$

Then

$$A = L + D + U, \tag{1.4.1}$$

with

$$L = \begin{bmatrix} 0 & & & \\ l_{11} & 0 & & \\ \vdots & \vdots & \ddots & \\ \dots & \dots & l_{n,n-1} & 0 \end{bmatrix},$$

$$D = \begin{bmatrix} d_{11} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & d_{nn} \end{bmatrix},$$

$$U = \begin{bmatrix} 0 & u_{12} & \dots & \dots \\ & 0 & \ddots & \vdots \\ & & 0 & u_{n-1,n} \\ & & & 0 \end{bmatrix}$$

and $A\mathbf{x} = \mathbf{b}$ becomes

$$(L + D + U)\mathbf{x} = \mathbf{b} \tag{1.4.2}$$

or

$$D\mathbf{x} = -(L + U)\mathbf{x} + \mathbf{b}. \tag{1.4.3}$$

Assuming that D^{-1} exists, this leads to the iterative scheme

$$\mathbf{x}^{(r+1)} = -D^{-1}(L + U)\mathbf{x}^{(r)} + D^{-1}\mathbf{b}, \quad x^{(0)} \text{ given}, \tag{1.4.4}$$

which is the *Jacobi method.*

The matrix formulation used in the above manipulations do not provide the easiest method of implementation. The obvious implementation is to rewrite the linear equations with the diagonal term on the left-hand side and then iterate equation by equation as in the example below. Consider the set of equations

$$\begin{bmatrix} 2 & 1 & 1 \\ -1 & 3 & 1 \\ 1 & 2 & -4 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 4 \\ -5 \\ 6 \end{bmatrix}$$

which yields the iteration

$$\begin{aligned} 2x_1^{(r+1)} &= 4 - x_2^{(r)} - x_3^{(r)} \\ 3x_2^{(r+1)} &= -5 + x_1^{(r)} - x_3^{(r)} \\ -4x_3^{(r+1)} &= 6 - x_1^{(r)} - 2x_2^{(r)}. \end{aligned}$$

Starting with the zero vector for $\mathbf{x}^{(0)}$ gives the results in Table 1.1.

Table 1.1.

r	x_1	x_2	x_3
0	0.0	0.0	0.0
1	2.000	−1.6667	−1.5000
2	3.5833	−0.5000	−1.83333
3	3.16667	0.13889	−0.85417
4	2.35764	−0.32639	−0.63889
5	2.48264	−0.66782	−1.07378
6	2.87080	−0.48119	−1.21325
7	2.84722	−0.30531	−1.02289
8	2.66410	−0.37663	−0.94085
9	2.65874	−0.46501	−1.02228
	⋮		⋮
20	2.72244	−0.41584	−1.02716

Intuitively, faster convergence would be expected the greater in magnitude the diagonal elements are compared with the off-diagonal elements. This is known as *diagonal dominance* and is only weakly exhibited in this example. Hence there is quite slow convergence to the correct result of (2.72222, −0.4166667, −1.0277778).

An improvement to this method is obtained by using the newly calculated elements as soon as they are available. Hence the elements which multiply L in (1.4.4) are known and could also be placed on the left-hand side of the iteration to give

$$(L + D)\mathbf{x}^{r+1} = -U\mathbf{x}^r + \mathbf{b}, \tag{1.4.5}$$

giving the *Gauss–Seidel iteration*

$$\mathbf{x}^{r+1} = -(L + D)^{-1}U\mathbf{x}^r + (L + D)^{-1}\mathbf{b}. \tag{1.4.6}$$

Note that (1.4.5) can be written

$$D\mathbf{x}^{r+1} = -L\mathbf{x}^{r+1} - U\mathbf{x}^r + \mathbf{b} \tag{1.4.7}$$

or

$$\mathbf{x}^{r+1} = \mathbf{x}^r + D^{-1}(\mathbf{b} - L\mathbf{x}^{r+1} - U\mathbf{x}^r - D\mathbf{x}^r). \tag{1.4.8}$$

Hence the new approximation is given by the old approximation together with a displacement (or correction).

In a practical form the above example set up for Gauss–Seidel iteration has the form:

$$\begin{aligned} 2x_1^{(r+1)} &= 4 - x_2^{(r)} - x_3^{(r)} \\ 3x_2^{(r+1)} &= -5 + x_1^{(r+1)} - x_3^{(r)} \\ -4x_3^{(r+1)} &= 6 - x_1^{(r+1)} - 2x_2^{(r+1)} \end{aligned}$$

and the iterations are shown in Table 1.2.

Table 1.2.

r	x_1	x_2	x_3
0	0.0	0.0	0.0
1	2.000	−1.000	−1.5000
2	3.25	−0.08333	−0.72917
3	2.40625	−0.62153	1.2092
4	2.91536	−0.21918	−0.91706
5	2.60444	−0.49283	−1.09531
6	2.79406	−0.37021	−0.98659
7	2.67840	−0.44501	−1.05290
8	2.74895	−0.39938	−1.01245
9	2.70592	−0.42721	−1.03712
	⋮		⋮
20	2.72229	−0.41662	−1.027737

Here the better convergence of the Gauss–Seidel method over Jacobi's method can be seen. In the practical applications of these methods to partial differential equations, the nature of the finite differencing often yields diagonally dominant matrices which give quite rapid convergence without the need for storing full matrices. Methods of increasing the convergence rate are of considerable interest and include the following approach.

If successive displacements are all one-signed, as they usually are for the approximating difference equations of elliptic problems, it would seem reasonable to expect convergence to be accelerated if a larger (displacement correction) was given than is defined above. This leads to the *successive over relaxation* or *SOR iteration* defined by

$$\mathbf{x}^{r+1} = \mathbf{x}^r + \omega D^{-1}[\mathbf{b} - L\mathbf{x}^{r+1} - U\mathbf{x}^r - D\mathbf{x}^r] \tag{1.4.9}$$

where ω, the acceleration parameter or relaxation factor, generally lies in the range $1 < \omega < 2$.

Thus, (1.4.9) becomes

$$(D + \omega L)\mathbf{x}^{r+1} = D\mathbf{x}^r + \omega\mathbf{b} - \omega U\mathbf{x}^r - \omega D\mathbf{x}^r \tag{1.4.10}$$

which can be rewritten as

$$\mathbf{x}^{r+1} = (D + \omega L)^{-1}[(1-\omega)D - \omega U]\mathbf{x}^r + \omega(D + \omega L)^{-1}\mathbf{b}. \tag{1.4.11}$$

All the iterative schemes described so far have the form

$$\mathbf{x}^{r+1} = G\mathbf{x}^r + H\mathbf{b}. \tag{1.4.12}$$

The rate of convergence of a scheme will be shown to be dictated by the magnitude of the dominant eigenvalue of the matrix G. Choosing the relaxation parameter in SOR suitably can result in savings in computational effort by a significant factor. In some special cases, optimal parameters can be found analytically, and the reader is referred to Young (1971) and Varga (1962) for further details. Usually experimentation and experience enable the user to obtain near optimal parameters.

The success of an iterative method depends on the rate of convergence. A point iterative method is one in which each component of $\mathbf{x}^r$ is calculated explicitly in terms of existing estimates of other components. A stationary iterative method is one in which $\mathbf{x}^r$ is calculated from known approximations by the same cycle of operations for all r. Jacobi, Gauss–Seidel and SOR are stationary iterative methods and have the form (1.4.12) where G is the iteration matrix and $H\mathbf{b}$ is a column of vectors of known values. Equation (1.4.12) was derived from the original equations $A\mathbf{x} = \mathbf{b}$ and hence the unique solution of $A\mathbf{x} = \mathbf{b}$ is the solution of

$$\mathbf{x} = G\mathbf{x} + H\mathbf{b}. \tag{1.4.13}$$

Define

$$\boldsymbol{\Delta}^r = \mathbf{x}^{r+1} - \mathbf{x}^r \quad \text{and} \quad \mathbf{e}^r = \mathbf{x} - \mathbf{x}^r \tag{1.4.14}$$

which leads recursively to

$$\mathbf{e}^r = G^r\mathbf{e}^0. \tag{1.4.15}$$

Hence the iteration will converge if and only if $\lim_{r\to\infty} G^r = 0$. We assume that the eigenvalues of G are real and that an eigenvector basis exists. Taking the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_N$ to be arranged so that $\mathbf{v}_i$ has corresponding eigenvalue λ_i, where

$$|\lambda_i| \leq |\lambda_{i-1}|, \quad i = 2, \ldots, n,$$

we note that $\mathbf{e}^0$ can be expressed uniquely as a linear combination of the eigenvectors to give

$$\mathbf{e}^0 = \gamma_1\mathbf{v}_1 + \gamma_2\mathbf{v}_2 + \cdots + \gamma_n\mathbf{v}_n \tag{1.4.16}$$

where γ_i, $i = 1, \ldots, n$, are scalars. Then

$$\begin{aligned} G\mathbf{e}^0 &= G(\gamma_1\mathbf{v}_1 + \gamma_2\mathbf{v}_2 + \cdots + \gamma_n\mathbf{v}_n) \\ &= \gamma_1\lambda_1\mathbf{v}_1 + \gamma_2\lambda_2\mathbf{v}_2 + \cdots + \gamma_n\lambda_n\mathbf{v}_n \\ &= \lambda_1\left(\gamma_1\mathbf{v}_1 + \gamma_2\frac{\lambda_2}{\lambda_1}\mathbf{v}_2 + \cdots + \gamma_n\frac{\lambda_n}{\lambda_1}\mathbf{v}_n\right) \end{aligned}$$

and hence

$$G^r\mathbf{e}^0 = \lambda_1^r\left(\gamma_1\mathbf{v}_1 + \gamma_2\left(\frac{\lambda_2}{\lambda_1}\right)^r\mathbf{v}_2 + \cdots + \gamma_n\left(\frac{\lambda_n}{\lambda_1}\right)^r\mathbf{v}_n\right). \tag{1.4.17}$$

Letting $\rho = |\lambda_1|$, results in the definitions that λ_1 is called the dominant eigenvalue of G and ρ is the spectral radius. Hence the iteration will converge for arbitrary $\mathbf{x}^0$ if and only if the spectral radius ρ of G is less than one. If r is large, then (1.4.17) can be written

$$\mathbf{e}^r \simeq \lambda_1^r\gamma_1\mathbf{v}_1.$$

Hence if the ith component in $\mathbf{e}^r$ is e_i^r and the ith component of $\mathbf{v}_1$ is v_{1i} then

$$|e_i^r| \simeq \rho^r|\gamma_1 v_{1i}|.$$

Ultimately, therefore the error in the approximation decreases by a factor $\sim 1/\rho$ with each iteration

$$\left(\frac{|e_i^r|}{|e_i^{r+1}|} \simeq \frac{1}{\rho}\right).$$

Suppose that we require to continue the iteration until no component of $\mathbf{e}^r$ exceeds E. We then require $|e_i^r| < E$, $i = 1, \ldots, n$. Set

$$m = \max_{1\le i\le n}|\gamma_1 v_{1i}|.$$

Then approximately, the requirement is $\rho^r m \le E$, or

$$r \ge \frac{\ln(m/E)}{-\ln\rho} \quad (\text{since } \rho < 1 \text{ and } -\ln\rho > 0). \tag{1.4.18}$$

Thus r, the number of iterations required to reduce to E the error in each component of $\mathbf{x}^r$ is inversely proportional to $-\ln\rho$.

How do we know when to terminate the iteration?

Realistically, our only measure is to test

$$\boldsymbol{\Delta}^r = \mathbf{x}^{r+1} - \mathbf{x}^r.$$

Now

$$\mathbf{x} \simeq \mathbf{x}^r + \boldsymbol{\Delta}^r + \boldsymbol{\Delta}^{r+1} + \boldsymbol{\Delta}^{r+2} + \cdots$$

and

$$\mathbf{e}^r \simeq \lambda_1\mathbf{e}^{r-1} \Rightarrow \mathbf{e}^{r+1} - \mathbf{e}^r \simeq \lambda_1(\mathbf{e}^r - \mathbf{e}^{r-1})$$

or

$$\mathbf{x}^{r+1} - \mathbf{x}^r \simeq \lambda_1(\mathbf{x}^r - \mathbf{x}^{r-1}) \Rightarrow \boldsymbol{\Delta}^r \simeq \boldsymbol{\Delta}^{r-1}.$$

Thus, for sufficiently large r,

$$\mathbf{x} \simeq \mathbf{x}^r + \boldsymbol{\Delta}^r(1 + \lambda_1 + \lambda_1^2 + \cdots) = \mathbf{x}^r + \frac{\boldsymbol{\Delta}^r}{1-\lambda_1}. \tag{1.4.19}$$

It follows that if we are to expect errors no greater than E in the components of $\mathbf{x}^r$ we must continue our iterations until

$$\max_{1\leq i\leq n}\left|\frac{\Delta_i^r}{1-\lambda_1}\right| < E. \tag{1.4.20}$$

This result tells us that the current correction $\boldsymbol{\Delta}^r$ should really be multiplied by $(1-\lambda_1)^{-1}$. This is an important result because if we want an approximation to $\mathbf{x}$ with an error no greater than E we might be tempted to terminate the iteration at the first r for which $||\boldsymbol{\Delta}^r|| \leq E$, where $||\,.\,||$ denotes the infinity norm $||\boldsymbol{\Delta}|| = \max_{1\leq i\leq n} |\Delta_i|$.

If $\lambda_1 = 0.99$ (which is quite possible), such a termination would give a very poor result, the iteration should be continued until $||\boldsymbol{\Delta}^r|| \simeq 0.01E$.

For most problems λ_1 will not be known analytically, in which case its value must be estimated. One straightforward way of doing this is as follows.

For sufficiently large r

$$\boldsymbol{\Delta}^r \simeq \lambda_1 \boldsymbol{\Delta}^{r-1}. \tag{1.4.21}$$

Hence

$$||\boldsymbol{\Delta}^r|| \simeq |\lambda_1|\,||\boldsymbol{\Delta}^{r-1}||,$$

so

$$|\lambda_1| = \rho \simeq \frac{||\boldsymbol{\Delta}^r||}{||\boldsymbol{\Delta}^{r-1}||}$$

where $||\boldsymbol{\Delta}^r||$ can be defined as

$$||\boldsymbol{\Delta}^r|| = \max_i |x_i^{r+1} - x_i^r|$$

or

$$||\boldsymbol{\Delta}^r|| = |x_1^{r+1} - x_1^r| + |x_2^{r+1} - x_2^r| + \cdots + |x_n^{r+1} - x_n^r|$$

or

$$||\boldsymbol{\Delta}^r|| = \left[(x_1^{r+1} - x_1^r)^2 + (x_2^{r+1} - x_2^r)^2 + \cdots + (x_n^{r+1} - x_n^r)^2\right]^{\frac{1}{2}}.$$

Equation (1.4.21) justifies the basis of the SOR iterative method because it proves that when λ_1 is positive the corresponding components of successive correction or displacement vectors are of the same sign. The following set of exercises may now be attempted on iterative methods of solution.

EXERCISES

1.10 Use both Jacobi's iterative method and that of Gauss–Seidel to find iteratively the solution of the linear equations

$$\begin{bmatrix} 2 & 1 & 0.5 \\ -1 & 3 & 1 \\ 0.5 & -1 & 4 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}.$$

1.11 Find the spectral radii of the G matrices of equation (1.4.12) for the matrix in Exercise 1.10 and hence find the theoretical rates of convergence. How do these rates compare with the actual rates obtained in the first exercise?

1.12 Now attempt the same problem with the SOR method with a range of ω from $1 < \omega < 1.4$. For this you will probably need to program the algorithm. Draw a graph of the rate of convergence against the relaxation parameter ω.

1.13 Investigate the rates of convergence of Jacobi's method and the Gauss–Seidel method on the matrix

$$A = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 1 & 2 & -3 \end{bmatrix}.$$

This is a pathological example: normally the Gauss–Seidel method is more rapidly convergent than Jacobi.

1.14 The SOR method for tridiagonal matrices has an optimum ω given by

$$\omega = \frac{2}{1 + \sqrt{1 - \rho(C_J)}}$$

where

$$C_J = D^{-1}(L + U)$$

is the iteration matrix for Jacobi's method. Apply this optimised method to the set of equations:

$$\begin{bmatrix} 2 & -2 & 0 & 0 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 6 & -1 \\ 0 & 0 & -1 & 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Experiment with values of ω slightly away from the optimum to show the sensitivity of the convergence rate to the ω value used.

1.15 Consider the matrix

$$\begin{bmatrix} 2 & -2 & 0 & 1 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 6 & -1 \\ 0 & 0 & -1 & 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

which differs from the one in Exercise 1.14 by just the element (1,4). Apply the SOR iteration to this matrix to see how much the change of one element affects the optimum ω. The new matrix is not tridiagonal so the theorem of Exercise 1.14 does not apply.

1.5 Further Results on Eigenvalues and Eigenvectors

In this section various results and proofs concerning eigenvalues and eigenvectors are collected together. These results are used freely in the following chapters. Let a square matrix A have eigenvector $\mathbf{x}$ and corresponding eigenvalue λ, then $A\mathbf{x} = \lambda\mathbf{x}$.

Hence

$$A(A\mathbf{x}) = A^2\mathbf{x} = \lambda A\mathbf{x} = \lambda^2\mathbf{x} \tag{1.5.1}$$

resulting in A^2 having eigenvalue λ^2 and eigenvector $\mathbf{x}$. Similarly

$$A^p\mathbf{x} = \lambda^p\mathbf{x}, \quad p = 3, 4, \ldots \tag{1.5.2}$$

and A^p has eigenvalue λ^p and eigenvector $\mathbf{x}$. These results may be generalised by defining

$$f(A) = a_pA^p + a_{p-1}A^{p-1} + \cdots + a_0I.$$

This is a polynomial in A when $a_p, \ldots, a_0$ are scalars. Then,

$$f(A)\mathbf{x} = (a_p\lambda^p + \cdots + a_0)\mathbf{x} = f(\lambda)\mathbf{x} \tag{1.5.3}$$

and $f(A)$ has eigenvalue $f(\lambda)$ and eigenvector $\mathbf{x}$. More generally we have the following simple theorem.

Theorem 1.4

The eigenvalue of $[f_1(A)]^{-1}f_2(A)$ corresponding to the eigenvector $\mathbf{x}$ is $f_2(\lambda)/f_1(\lambda)$, where $f_1(A)$ and $f_2(A)$ are polynomials in A.

Proof

We have

$$f_1(A)\mathbf{x} = f_1(\lambda)\mathbf{x}, \quad f_2(A)\mathbf{x} = f_2(\lambda)\mathbf{x}$$

Pre-multiply by $[f_1(A)]^{-1}$ to give

$$[f_1(A)]^{-1}[f_1(A)]\mathbf{x} = [f_1(A)]^{-1}f_1(\lambda)\mathbf{x}$$

and hence

$$[f_1(A)]^{-1}\mathbf{x} = [f_1(\lambda)]^{-1}\mathbf{x}$$

and

$$[f_1(A)]^{-1}f_2(A)\mathbf{x} = f_2(\lambda)[f_1(A)]^{-1}\mathbf{x}.$$

Eliminating $[f_1(A)]\mathbf{x}$ gives

$$[f_1(A)]^{-1}f_2(A)\mathbf{x} = \frac{f_2(\lambda)}{f_1(\lambda)}\mathbf{x}.$$

Similarly the eigenvalue of $f_2(A)[f_1(A)]^{-1}$ corresponding to the eigenvector $\mathbf{x}$ is $f_2(\lambda)/f_1(\lambda)$.

The second set of results concerns the eigenvalues of an order n tridiagonal matrix and forms the next theorem.

Theorem 1.5

The eigenvalues of the order n tridiagonal matrix

$$\begin{pmatrix} a & b & & & & \\ c & a & b & & & \\ & c & a & b & & \\ & \ddots & \ddots & \ddots & & \\ & & & c & a & b \\ & & & & c & a \end{pmatrix}$$

are

$$\lambda_s = a + 2\left[\sqrt{bc}\right] \cos \frac{s\pi}{n+1}, \quad s = 1(1)n \tag{1.5.4}$$

where a, b and c may be real or complex. This class of matrices arises commonly in the study of stability of the finite difference processes, and a knowledge of its eigenvalues leads immediately into useful stability conditions.

Proof

Let λ represent an eigenvalue of A and $\mathbf{v}$ the corresponding eigenvector with components $v_1, v_2, \ldots, v_n$. Then the eigenvalue equation $A\mathbf{v} = \lambda\mathbf{v}$ gives

$$\begin{aligned} (a-\lambda)v_1 + bv_2 &= 0 \\ cv_1 + (a-\lambda)v_2 + bv_3 &= 0 \\ &\vdots \\ cv_{j-1} + (a-\lambda)v_j + bv_{j+1} &= 0 \\ &\vdots \\ cv_{n-1} + (a-\lambda)v_n &= 0. \end{aligned}$$

Now define $v_0 = v_{n+1} = 0$ and these n equations can be combined into one difference equation

$$cv_{j-1} + (a-\lambda)v_j + bv_{j+1} = 0, \quad j = 1, \ldots, n. \tag{1.5.5}$$

The solution is of the form $v_j = Bm_1^j + Cm_2^j$ where B and C are arbitrary constants and m_1, m_2 are roots of the equation

$$C + (a-\lambda)m + bm^2 = 0. \tag{1.5.6}$$

Hence the conditions

$$v_0 = v_{n+1} = 0$$

give

$$0 = B + C, \quad \text{and} \quad 0 = Bm_1^{n+1} + Cm_2^{n+1}$$

which implies

$$\left(\frac{m_1}{m_2}\right)^{n+1} = 1 = e^{i2s\pi}, \quad s = 1, 2, \ldots, n,$$

or

$$\frac{m_1}{m_2} = e^{\frac{i2s\pi}{n+1}}.$$

From (1.5.6), $m_1 m_2 = c/b$ and $m_1 + m_2 = -(a - \lambda)/b$. Hence

$$m_1 = \left(\frac{c}{b}\right)^{\frac{1}{2}} e^{\frac{is\pi}{n+1}} \quad \text{and} \quad m_2 = \left(\frac{c}{b}\right)^{\frac{1}{2}} e^{\frac{-is\pi}{n+1}}$$

which gives

$$\lambda = a + b(m_1 + m_2) = a + b\left(\frac{c}{b}\right)^{\frac{1}{2}} \left(e^{\frac{is\pi}{n+1}} + e^{\frac{-is\pi}{n+1}}\right).$$

Hence the n eigenvalues are

$$\lambda_s = a + 2\sqrt{bc}\cos\frac{s\pi}{n+1}, \quad s = 1, 2, \ldots, n \tag{1.5.7}$$

as required.

The jth component of the corresponding eigenvector is

$$\begin{aligned} v_j &= Bm_1^j + Cm_2^j \\ &= B\left(\frac{c}{b}\right)^{\frac{j}{2}} \left(e^{\frac{ijs\pi}{n+1}} - e^{\frac{-ijs\pi}{n+1}}\right) \\ &= 2iB\left(\frac{c}{b}\right)^{\frac{j}{2}} \sin\left(\frac{js\pi}{n+1}\right). \end{aligned}$$

So the eigenvector $\mathbf{v}_s$ corresponding to λ_s is

$$\begin{aligned} \mathbf{v}_s^T = \Bigg[&\left(\frac{c}{b}\right)^{\frac{1}{2}} \sin\frac{s\pi}{n+1}, \left(\frac{c}{b}\right) \sin\frac{2s\pi}{n+1}, \\ &\left(\frac{c}{b}\right)^{\frac{3}{2}} \sin\frac{3s\pi}{n+1}, \ldots, \left(\frac{c}{b}\right)^{\frac{n}{2}} \sin\frac{ns\pi}{n+1}\Bigg]. \end{aligned} \tag{1.5.8}$$

As an example consider the tridiagonal matrix

$$\begin{pmatrix} 1-2r & r & & & & \\ r & 1-2r & r & & & \\ & r & 1-2r & r & & \\ & & \ddots & \ddots & \ddots & \\ & & & r & 1-2r & r \\ & & & & r & 1-2r \end{pmatrix}$$

of order $n-1$ with

$$a = 1-2r, \quad b = r, \quad c = r.$$

Then the previous theorem tells us that the eigenvalues are

$$\begin{aligned}\lambda_s &= (1-2r) + 2r\left(\frac{r}{r}\right)^{\frac{1}{2}} \cos\frac{s\pi}{n} \\ &= 1 - 2r\left[1 - \cos\frac{s\pi}{n}\right] \\ &= 1 - 4r\sin^2\frac{s\pi}{2n}.\end{aligned}$$

Many of the methods which arise in the solution of partial differential equations require the solution of a tridiagonal set of linear equations, and for this special case the usual elimination routine can be simplified. The algorithm which results is called the Thomas algorithm for tridiagonal systems, and is described below.

Suppose that it is required to solve

$$\begin{pmatrix} b_1 & -c_1 & & & \\ -a_2 & b_2 & -c_2 & & \\ & -a_3 & b_3 & -c_3 & \\ & \ddots & \ddots & \ddots & \\ & & -a_{n-1} & b_{n-1} & -c_{n-1} \\ & & & -a_n & b_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}.$$

The algorithm is based on Gauss elimination. In each column only one sub-diagonal element is to be removed. In each equation b_i and d_i, $i = 2, \ldots, n$, change as a result of the elimination. Denote the quantities that replace b_i and d_i by α_i and s_i respectively. For convenience set $\alpha_1 = b_1$ and $s_1 = d_1$ then

$$\begin{aligned}\alpha_2 &= b_2 - \frac{c_1 a_2}{\alpha_1}, \quad s_2 = d_2 + \frac{s_1 a_2}{\alpha_1}, \\ \alpha_3 &= b_3 - \frac{c_2 a_3}{\alpha_2}, \quad s_3 = d_3 + \frac{s_2 a_3}{\alpha_2},\end{aligned}$$

etc.

In general

$$\alpha_i = b_i - \frac{c_{i-1}a_i}{\alpha_{i-1}}, \quad s_i = d_i + \frac{s_{i-1}a_i}{\alpha_{i-1}}. \tag{1.5.9}$$

Once the elimination is complete the x_i, $i = 1, \ldots, n$, are found recursively by back substitution.

The complete algorithm may be expressed as:

$$\begin{aligned}\alpha_1 &= b_1, \qquad s_1 = d_1, \\ \alpha_i &= b_i - \frac{c_{i-1}a_i}{\alpha_{i-1}}, \qquad s_i = d_i - \frac{s_{i-1}a_i}{\alpha_{i-1}}, \quad i = 2, \ldots, n, \\ x_n &= \frac{s_n}{\alpha_n}, \qquad x_i = \frac{(s_i + c_i x_{i+1})}{\alpha_i}, \quad i = (n-2), \ldots, 1.\end{aligned}$$

Conditions for the applicability of the method are considered next.

We have not used partial pivoting and so we need to investigate the conditions for which the multipliers a_i/α_{i-1}, $i = 2,\ldots,n$, have magnitude not exceeding unity for stable forward elimination and c_i/α_i, $i = 2,\ldots,n-1$, have magnitude not exceeding unity for stable back substitution.

Suppose that $a_i > 0$, $b_i > 0$, $c_i > 0$ then,

(i) assuming that $b_i > a_{i+1} + c_{i-1}$, $i = 1,\ldots,n-1$, the forward elimination is stable; and

(ii) assuming that $b_i > a_i + c_i$, $i = 1,\ldots,n-1$, the back-substitution is stable.

The proof can be found in Smith (1978).

Some assorted exercises on these ideas are now presented.

EXERCISES

1.16 Use the characteristic polynomial directly to confirm that the eigenvalues given in (1.5.4) are correct for $n = 2$.

1.17 Find the characteristic polynomial and hence the eigenvalues of the matrix

$$\begin{bmatrix} 4 & 1 & 0 \\ 2 & 4 & 1 \\ 0 & 2 & 4 \end{bmatrix}$$

and compare the result with the formula (1.5.4).

1.18 Use the Thomas algorithm to solve the tridiagonal set of equations

$$\begin{bmatrix} 4 & 1 & 0 \\ 2 & 4 & 1 \\ 0 & 2 & 4 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

1.19 By counting operations establish that Gaussian elimination requires the order of $n^3/3$ multiplication and division operations. This is a measure of the work load in the algorithm. The easiest way to establish this result is to code up the algorithm (which will be a useful tool for later anyway) and then use the formulae:

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2},$$

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6},$$

$$\sum_{i=1}^{n} i^3 = \left[\frac{n(n+1)}{2}\right]^2.$$

What is the equivalent count for Thomas's algorithm?

1.20 Compare the work load in Thomas' algorithm with that for say m iterations of Gauss–Seidel. Given the convergence rate from the eigenvalues of the G matrix of (1.4.12), construct advice for prospective users on whether to use the Thomas algorithm or Gauss–Seidel.

1.21 Extend the Thomas algorithm to deal with upper Hessenberg matrices with the form

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & \dots & a_{1n} \\ b_2 & a_{22} & a_{23} & \dots & \dots & a_{2n} \\ 0 & b_3 & a_{33} & \dots & \dots & a_{3n} \\ \vdots & & & & & \vdots \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \dots & b_n & a_{nn} \end{bmatrix}$$

which is tridiagonal with non-zero elements in the top right-hand part of the matrix.

1.22 Extend Thomas's algorithm to quindiagonal matrices which have in general diagonal elements with two non-zero elements on either side in each row, except in the first two and last two rows which just have two non-zero elements on one side for the first row, and in addition one non-zero element on the opposite side in the second row.

1.6 Classification of Second Order Partial Differential Equations

Consider a general second order quasi-linear equation defined by the equation

$$Rr + Ss + Tt = W \tag{1.6.1}$$

where

$$p = \frac{\partial z}{\partial x}, \quad q = \frac{\partial z}{\partial y}, \quad r = \frac{\partial^2 z}{\partial x^2}, \quad s = \frac{\partial^2 z}{\partial x \partial y} \quad \text{and} \quad t = \frac{\partial^2 z}{\partial y^2} \tag{1.6.2}$$

with

$$R = R(x, y), \quad S = S(x, y), \quad T = T(x, y) \quad \text{and} \quad W = W(x, y, z, p, q). \tag{1.6.3}$$

Then the characteristic curves for this equation are defined as curves along which highest partial derivatives are not uniquely defined. In this case these

derivatives are the second order derivatives r, s and t. The set of linear algebraic equations which these derivatives satisfy can be written down in terms of differentials, and the condition for this set of linear equations to have a non-unique solution will yield the equations of the characteristics, whose significance will then become more apparent. Hence the linear equations follow as $dz = pdx + qdy$ and also

$$\begin{aligned} dp &= rdx + sdy \\ dq &= sdx + tdy \end{aligned} \tag{1.6.4}$$

to give the linear equations

$$\left.\begin{aligned} Rr + Ss + Tt &= W \\ rdx + sdy &= dp \\ sdx + tdy &= dq \end{aligned}\right\} \tag{1.6.5}$$

and there will be no unique solution when

$$\begin{vmatrix} R & S & T \\ dx & dy & 0 \\ 0 & dx & dy \end{vmatrix} = 0 \tag{1.6.6}$$

which expands to give the differential equation

$$R\left(\frac{dy}{dx}\right)^2 - S\left(\frac{dy}{dx}\right) + T = 0. \tag{1.6.7}$$

But when the determinant in (1.6.6) is zero, the other determinants in Cramer's rule for the solution of (1.6.5) will also be zero, for we assume that (1.6.5) does not have a unique solution. Hence the condition

$$\begin{vmatrix} R & T & W \\ dx & 0 & dp \\ 0 & dy & dq \end{vmatrix} = 0 \tag{1.6.8}$$

also holds, and gives an equation which holds along a characteristic, namely

$$-Rdy\ dp - T\ dx\ dq + W\ dxdy = 0 \tag{1.6.9}$$

or

$$R\frac{dp}{dx}\frac{dy}{dx} + T\frac{dq}{dx} - W\frac{dy}{dx} = 0. \tag{1.6.10}$$

Returning now to (1.6.6), this equation is a quadratic in dy/dx and there are three possible cases which arise. If the roots are real the characteristics form two families of real curves. A partial differential equation resulting in real characteristics is said to be hyperbolic. The condition is that

$$S^2 - 4RT > 0. \tag{1.6.11}$$

The second case is when the roots are equal to give the parabolic case and the condition

$$S^2 - 4RT = 0, \tag{1.6.12}$$

and when the roots are complex the underlying equation is said to be elliptic with the condition

$$S^2 - 4RT < 0. \tag{1.6.13}$$

The importance of characteristics only becomes apparent at this stage. The first feature is the use of characteristics to classify equations. The methods that will be used subsequently are quite different from type to type. In the case of hyperbolic equations, the characteristics are real and are used directly in the solution. Characteristics also play a role in reducing equations to a standard or canonical form. Consider the operator

$$R\frac{\partial^2}{\partial x^2} + S\frac{\partial^2}{\partial x \partial y} + T\frac{\partial^2}{\partial y^2} \tag{1.6.14}$$

and put $\xi = \xi(x, y)$, $\eta = \eta(x, y)$ and $z = \zeta$ to see what a general change of variable yields. The result is the operator

$$\begin{aligned} A(\xi_x, \xi_y)\frac{\partial^2 \zeta}{\partial \xi^2} + 2B(\xi_x, \xi_y, \eta_x, \eta_y)\frac{\partial^2 \zeta}{\partial \xi \partial \eta} \\ + A(\eta_x, \eta_y)\frac{\partial^2 \zeta}{\partial \eta^2} = F(\xi, \eta, \zeta, \zeta_\zeta, \zeta_\eta) \end{aligned} \tag{1.6.15}$$

where

$$A(u, \nu) = Ru^2 + Su\nu + T\nu^2 \tag{1.6.16}$$

and

$$B(u_1, \nu_1, u_2, \nu_2) = Ru_1u_2 + \frac{1}{2}S(u_1\nu_2 + u_2\nu_1) + T\nu_2\nu_2. \tag{1.6.17}$$

The question is now asked for what ξ and η do we get the simplest form? Certainly if ξ and η can be found to make the coefficients A equal to zero, then a simplified form will result. However the condition that A should be zero is a partial differential equation of first order which can be solved analytically (Sneddon, 1957). Different cases arise in the three classifications. In the hyberbolic case when $S^2 - 4RT > 0$, let $R\alpha^2 + S\alpha + T = 0$ have roots λ_1 and λ_2 then $\xi = f_1(x, y)$ and $\eta = f_2(x, y)$ where $f_1(x, y)$ and $f_2(x, y)$ are the solutions of the two factors in the related ordinary differential equations

$$\left[\frac{dy}{dx} + \lambda_1(x, y)\right]\left[\frac{dy}{dx} + \lambda_2(x, y)\right] = 0. \tag{1.6.18}$$

Hence the required transformations are precisely the defining functions of the characteristic curves. It follows that with this change of variable the partial differential equation becomes

$$\frac{\partial^2 \zeta}{\partial \eta \partial \xi} = \phi(\xi, \eta, \zeta, \zeta_\zeta, \zeta_\eta) \tag{1.6.19}$$

which is the canonical form for the hyperbolic case.

In the parabolic case, $S^2 - 4RT = 0$, there is now only one root, and any independent function is used for the other variable in the transformation. Hence $A(\xi_x, \xi_y) = 0$, but it is easy to show in general that

$$A(\xi_x, \xi_y)A(\eta_x, \eta_y) - B^2(\xi_x, \xi_y, \eta_x, \eta_y) = (4RT - S^2)(\xi_x\eta_y - \xi_y\eta_x)^2$$

and therefore as $S^2 = 4RT$, we must have $B(\xi_x, \xi_y, \eta_x, \eta_y) = 0$ and $A(\eta_x, \eta_y) \neq 0$ as η is an independent function of x and y. Hence when $S^2 = 4RT$, the transformation $\xi = f_1(x, y)$ and $\eta =$ any independent function yields

$$\frac{\partial^2 \zeta}{\partial \eta^2} = \phi_1(\xi, \eta, \zeta, \zeta_\zeta, \zeta_\eta) \tag{1.6.20}$$

which is the canonical form for a parabolic equation.

In the elliptic case there are again two sets of characteristics but they are now complex. Writing $\xi = \alpha + i\beta$ and $\eta = \alpha - i\beta$ gives the real form

$$\frac{\partial^2 \zeta}{\partial \xi \partial \nu} = \frac{1}{4}\left(\frac{\partial^2 \zeta}{\partial \alpha^2} + \frac{\partial^2 \zeta}{\partial \beta^2}\right) \tag{1.6.21}$$

and hence the elliptic canonical form

$$\frac{\partial^2 \zeta}{\partial \alpha^2} + \frac{\partial^2 \zeta}{\partial \beta^2} = \psi(\alpha, \beta, \zeta, \zeta_\alpha, \zeta_\beta). \tag{1.6.22}$$

Note that Laplace's equation is in canonical form as is the heat equation, but the wave equation is not. As an example of reduction to canonical form consider the linear second order partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + 2\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} + c^2 \frac{\partial u}{\partial y} = 0. \tag{1.6.23}$$

Then the equation of the characteristic curves is

$$\left(\frac{dy}{dx}\right)^2 - 2\frac{dy}{dx} + 1 = 0 \tag{1.6.24}$$

or factorising

$$\left(\frac{dy}{dx} - 1\right)^2 = 0. \tag{1.6.25}$$

Therefore the transformation for the canonical form is:

$$\left.\begin{array}{rcl} p & = & x - y \\ q & = & x \end{array}\right\} \tag{1.6.26}$$

and the required partial derivatives are:

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial p^2}\left(\frac{\partial p}{\partial x}\right)^2 + \frac{\partial^2 u}{\partial q^2}\left(\frac{\partial q}{\partial x}\right)^2 \tag{1.6.27}$$

and

$$\frac{\partial^2 u}{\partial x \partial y} = \frac{\partial^2 u}{\partial p^2}\left(\frac{\partial p}{\partial y}\frac{\partial p}{\partial x}\right) \tag{1.6.28}$$

which yields the reduced form

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2} + 2\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} &= \frac{\partial^2 u}{\partial p^2} + \frac{\partial^2 u}{\partial q^2} - 2\frac{\partial^2 u}{\partial p^2} + \frac{\partial^2 u}{\partial p^2} \\ &= \frac{\partial^2 u}{\partial q^2}\end{aligned} \tag{1.6.29}$$

with the transformed equation being

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial q^2} = \frac{\partial u}{\partial p}. \tag{1.6.30}$$

From a numerical point of view, the canonical forms reduce the number of different types of equation for which solutions need to be found. Effectively effort can be concentrated on the canonical forms alone, though this is not always the best strategy, and in this spirit the parabolic type will now be considered in detail in the next chapter. Before considering this work the reader may wish to pursue some of the ideas of the previous section in the following exercises.

EXERCISES

1.23 Classify the following partial differential equations as parabolic, elliptic or hyperbolic:

(a) $\dfrac{\partial^2 \phi}{\partial x^2} + \dfrac{\partial^2 \phi}{\partial x \partial y} + \dfrac{\partial^2 \phi}{\partial y^2} = 0$

(b) $\dfrac{\partial^2 \phi}{\partial t^2} - \dfrac{\partial^2 \phi}{\partial x^2} + \dfrac{\partial \phi}{\partial x} = 0$

(c) $\dfrac{\partial \phi}{\partial t} - \dfrac{\partial^2 \phi}{\partial x^2} - \dfrac{\partial \phi}{\partial x} = 0$

(d) $\dfrac{\partial^2 \phi}{\partial x^2} + x\dfrac{\partial^2 \phi}{\partial y^2} = 0.$

1.24 Find the regions of parabolicity, ellipticity and hyperbolicity for the partial differential equation:

$$\frac{\partial^2 u}{\partial x^2} + 3x^2 y^2 \frac{\partial^2 u}{\partial x \partial y} + (x + y)\frac{\partial^2 u}{\partial y^2} = u$$

and sketch the resulting regions in the (x, y) plane.

1.25 Find the analytic form of the characteristic curves for the partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + 2\left(x + \frac{1}{y}\right)\frac{\partial^2 u}{\partial x \partial y} + \frac{4x}{y}\frac{\partial^2 u}{\partial y^2} = xy$$

and hence categorise the equation.

1.26 Reduce the equation

$$\frac{\partial^2 z}{\partial x^2} - 6\frac{\partial^2 z}{\partial x \partial y} + 9\frac{\partial^2 z}{\partial y^2} = \frac{\partial z}{\partial y}$$

to canonical form.

1.27 Reduce the equation

$$\frac{\partial^2 z}{\partial x^2} + 3\frac{\partial^2 z}{\partial x \partial y} + \frac{\partial^2 z}{\partial y^2} = 0$$

to canonical form, and hence find the general analytic solution.

1.28 Reduce the equation

$$\frac{\partial^2 z}{\partial x^2} + 2\frac{\partial^2 z}{\partial x \partial y} + 3\frac{\partial^2 z}{\partial y^2} = z$$

to canonical form. Make a further transformation to obtain a real canonical form.