Skript zur Numerischen Mathematik

Frank Stenger, Universität Utah

1996

This page intentionally left blank.

Dies Buch enthält Teile eines (offensichtlich unkorrigierten) Skripts zur Numerischen Mathematik. Die Seitenzahlen sind durcheinandergeraten; in jedem Kapitel wird neu mit Seitenzahl 1 begonnen. Ein Inhaltsverzeichnis existiert leider nicht. Das Buch enthält die Kapitel:

- 2. Lineare Gleichungen
- 3. Nichtlineare Gleichungen
- 4. Polynomapproximation
- 5. Numerische Integration
- 6. Asymptotische Methoden
- 7. Numerische Lösung gewöhnlicher Differentialgleichungen

Frank Wübbeling

This page intentionally left blank.

Chapter 2

LINEAR EQUATIONS and MATRIX INVERSION

2.1 Properties of Matrices

2.1.1 Definitions and Notation

Definition 2.1.1 *An $m \times n$ matrix A is a rectangular array of numbers denoted by $A = [a_{i,j}], 1 \leq i \leq m,\ 1 \leq j \leq n$, or less compactly by*

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}.$$

The element $a_{i,j}$ appearing in the i-th row and j-th column of A is called the (i,j)th element of A.

Definition 2.1.2 *If $m = n$ then A is said to be a square matrix of order n.*

Definition 2.1.3 *If $m = 1$ then A is a row vector of order n while if $n = 1$ then A is a column vector of order m.*

When we wish to emphasize that a matrix is a vector we use a lower case letter. Upper case letters denote a matrix that is either not a vector or not thought of as a vector. For square matrices we adhere to the custom of saying "matrix of order n" instead of "square matrix of order n". Further, if we wish to stress n we write A_n in place of A. With respect to vectors, the letters $a, b, c, \ldots, q$ denote column vectors whose elements are part of given data while the letters x, y, z stand for column vectors determined by solving equations of the form $Ax = f$.

Definition 2.1.4 *A^T denotes the transpose of a matrix A and is defined by $A^T = [a_{i,j}]^T = [a_{j,i}]$. If A is $m \times n$ then A^T is $n \times m$. If a is a column vector then a^T is a row vector.*

Definition 2.1.5 *The $m \times n$ matrices $A = [a_{i,j}]$ and $B = [b_{i,j}]$ are equal if and only if $a_{i,j} = b_{i,j}$ for all i, j.*

Definition 2.1.6 *The square matrix A is symmetric if $A^T = A$.*

Definition 2.1.7 *The system of equations*

$$Ax = f \tag{2.1.1}$$

is said to be consistent for the given A and f if there exists a vector x such that (2.1.1) is satisfied. Otherwise (2.1.1) is said to be inconsistent.

Please note that if m, the row dimension of A, is greater than n, the column dimension of A, then one should proceed under the assumption that the system (2.1.1) is inconsistent. Despite lacking a solution inconsistent systems of the form (2.1.1) often arise in practice. One could simply ignore enough equations to generate a solution but a better alternative is to choose the x that minimizes the average error of the m equations. This puts us into the realm of least squares analysis, a subject we wish to defer for a later chapter. Therefore, for purposes of this chapter we address solve (2.1.1) assuming that the row dimension of A is less than its column dimension

When A is a dense matrix, that is most of its elements are nonzero, then factoring A into the product of upper and lower triangular matrices is usually the most economical means of solving $Ax = f$.

Definition 2.1.8 *$L = [l_{i,j}], 1 \leq i, j \leq n$, is a lower triangular matrix if $l_{i,j} = 0$ when $i < j$. Similarly, $U = [u_{i,j}], 1 \leq i, j \leq n$, is an upper triangular matrix if $u_{i,j} = 0$ when $i > j$.*

Definition 2.1.9 *The square matrix A is diagonal if it is both lower and upper triangular.*

Definition 2.1.10 *The unit or identity matrix of order n is denoted by I where unambiguous and I_n otherwise. I is defined by $I = [\delta_{i,j}]$ where*

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{2.1.2}$$

Definition 2.1.11 *The determinant $d(A)$ of a square matrix A is defined by*

$$d(A) = \sum_{\sigma} \mu(\sigma) a_{1,\sigma(1)} a_{2,\sigma(2)} \cdots a_{n,\sigma(n)} \tag{2.1.3}$$

where σ runs over all $n!$ permutations of n integers, and $\mu(\sigma)$ is -1 if σ is an odd permuation, $+1$ if σ is an even permutation.

Since this definition may be unfamiliar to the student we offer the alternative definition given in terms of minors.

Definition 2.1.12 *The minor $M_{i,j}$ of the matrix A is formed by throwing away row i and column j of A.*

Definition 2.1.13 *The cofactor $A_{i,j}$ of A is*

$$A_{i,j} = (-1)^{i+j} d(M_{i,j})$$

Definition 2.1.14 *The determinant of the $n \times n$ matrix A is a linear combination of any row i with the cofactors of row i; that is,*

$$d(A) = a_{i,1}A_{i,1} + a_{i,2}A_{i,2} + \cdots + a_{i,n}A_{i,n} \ . \tag{2.1.4}$$

A familiar means of evaluating the determinant by hand uses (2.1.4) repeatedly along with

$$d\left(\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}\right) = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}$$

The identity matrix and the notion of determinants are necessary for defining eigenvalues.

Definition 2.1.15 *If for a nonzero vector x there exists a scalar λ such that*

$$Ax = \lambda I x = \lambda x$$

then λ, x is called an eigenvalue, eigenvector pair of A.

Note that for the eigenvector $x, \lambda I$ acting on x is equivalent to A acting on x.

Theorem 2.1.16 *The number λ is an eigenvalue of A if and only if $d(A - \lambda I) = 0$.*

Definition 2.1.17 *For the $n \times n$ matrix A the n-th degree polynomial*

$$p(\lambda) = d(A - \lambda I)$$

is called the characteristic equation of A.

Definition 2.1.18 *The set of all eigenvalues of the matrix A is called the spectrum of A.*

Definition 2.1.19 *The square matrix A is said to be invertible if there exists a matrix B such that $AB = BA = I$. B is traditionally denoted A^{-1}.*

We mention that it is possible to calculate A^{-1} in terms of cofactors and $d(A)$, however, numerically it is difficult to find a worse choice of algorithms for the task.

2.1.2 Matrix Arithmetic

Throughout the list of definitions it is clear that some notion of matrix arithmetic is necessary. The easiest operation is scalar multiplication where if λ is a scalar and $A = [a_{i,j}]$ then the elements of λA are $\lambda a_{i,j}$. The addition of matrices $A = [a_{i,j}]$ and $B = [b_{i,j}]$ requires that their dimensions be identical; if so $C = A + B, C = [c_{i,j}]$, is given elementwise by

$$c_{i,j} = a_{i,j} + b_{i,j} \ .$$

Matrix multiplication is considerably less intuitive. If the number of columns of $A = [a_{i,j}]$, $1 \le i \le m$, $1 \le j \le n$, equals the number of rows of $B = [b_{j,k}]$, $1 \le j \le n$, $1 \le k \le p$, then $C = AB, C = [c_{i,k}]$ is defined with

$$c_{i,k} = \sum_{j=1}^{n} a_{i,j} b_{j,k}, \ 1 \le i \le m, 1 \le k \le p.$$

Thus the row dimension of C is inherited from A and the column dimension from B.

Theorem 2.1.20 *Let α and β be scalars, then assuming compatible dimensions the following are easy to prove using properties of real numbers.*

(i) $A + B = B + A$

(ii) $(A + B) + C = A + (B + C)$

(iii) $\alpha(A + B) = \alpha A + \alpha B$

(iv) $(\alpha + \beta)A = \alpha A + \beta A$

(v) $A = IA$

(vi) $B = BI$

(vii) $A(B + C) = AB + AC$

(viii) $(A + B)C = AC + BC$

(xi) $(\alpha A)B = A(\alpha B)$

(x) $(AB)C = A(BC)$

(xi) $(A + B)^T = A^T + B^T$

(xii) $(A^T)^T = A$

(xiii) $(AB)^T = B^T A^T$.

Theorem 2.1.21 *Assuming the matrices A and B are both of order n, it can be shown that*

(xiv) $(A^{-1})^{-1} = A$, *if A is invertible*

(xv) $(AB)^{-1} = B^{-1}A^{-1}$, *if A and B are invertible*

(xvi) $d(AB) = d(A)d(B)$.

Triangular matrices lead to computational savings when performing matrix arithmetic. For instance, the determinant of a triangular matrix is the product of its diagonal entries (see Problem 2.1.9). We emphasize at this point that formula (2.1.3) is not a practical means of evaluating $d(A_n)$ for $n > 4$, rather it is usually more economical to first factor A_n into LU, then use $d(A_n) = d(L)d(U)$.

2.1.3 Vector Spaces and Linear Independence

In this subsection we address the existence and uniqueness of solutions to $Ax = f$.

Definition 2.1.22 *A set S is a vector space if for all $\alpha, \beta \in \mathbb{R}, v_1, v_2 \in S$, $\alpha v_1 + \beta v_2 \in S$.*

Example 2.1.23 The set of column vectors with n components is referred to as the vector space $\mathbb{R}^n$.

Example 2.1.24 The set of vectors $[x, y, 0]^T$ is a vector space since

$$\alpha \begin{bmatrix} x_1 \\ y_1 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} x_2 \\ y_2 \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha x_1 + \beta x_2 \\ \alpha y_1 + \beta y_2 \\ 0 \end{bmatrix}$$

maintains the 0 in the third position.

Example 2.1.25 The set of vectors $[x, y, 1]^T$ is not a vector space since for $\alpha = \beta = 1$

$$\alpha \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} + \beta \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ y_1 + y_2 \\ 2 \end{bmatrix}.$$

Definition 2.1.26 *A subset S_1 of a vector space S is called a subspace of S if S_1 is a vector space in its own right.*

Example 2.1.24 implies the x-y plane is a two-dimensional subspace of $\mathbb{R}^3$, that is, any linear combination of vectors in the x-y plane remains in the x-y plane. More generally, $\mathbb{R}^m$ is a subspace of $\mathbb{R}^n$ if $m \leq n$.

It is well-known that any vector $[x, y, z]^T$ in $\mathbb{R}^3$ may be written in terms of $\hat{\imath} = [1, 0, 0]^T$, $\hat{\jmath} = [0, 1, 0]^T$, and $\hat{k} = [0, 0, 1]^T$ as $[x, y, z]^T = x\hat{\imath} + y\hat{\jmath} + z\hat{k}$. Alternatively the vector may also be written

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \frac{1}{3}z \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix} + (\frac{1}{2}y - \frac{1}{6}z) \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + (x - \frac{1}{2}y + \frac{1}{6}z) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

However, using the vectors $[1, 2, 3]^T, [4, 5, 6]^T$, and $[7, 8, 9]^T$ it is not possible to represent $[x, y, z]^T = [1, 1, 0]^T$ while if

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = -\frac{1}{3} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \frac{1}{3} \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = -\frac{1}{3} \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} + \frac{1}{3} \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix}$$

then its representation is not unique. The criterion governing the situation is the linear independence of vectors.

Definition 2.1.27 *The set of vectors $v_1, v_2, \ldots, v_n$ is linearly independent if and only if*

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0 \Rightarrow \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0.$$

Otherwise $v_1, v_2, \ldots, v_n$ are linearly dependent.

Definition 2.1.28 *The span of $v_1, v_2, \ldots, v_n$ is the set of all linear combinations of the vectors.*

Definition 2.1.29 *If $v_1, v_2, \ldots, v_n$ are linearly independent then they are said to form a basis for their span.*

An immediate consequence is that if $\{v_1, v_2, \ldots, v_n\}$ are linearly independent and $f \in \text{span}\{v_1, v_2, \ldots, v_n\}$ then f admits a unique representation in terms of the basis.

Turning to $A_n x = f$, we first write the system as

$$\begin{bmatrix} & \vdots & & \vdots & & \vdots & \\ a_1 & \vdots & a_2 & \vdots & \cdots & \vdots & a_n \\ & \vdots & & \vdots & & \vdots & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = f \tag{2.1.5}$$

where a_j is the j-th column of A treated as a vector. Expanding (2.1.5) yields

$$x_1 a_1 + x_2 a_2 + \cdots + x_n a_n = f$$

or, in other words, solving the system is equivalent to finding the correct linear combination of $a_1, a_2, \ldots, a_n$ that produces f. If no such linear combination exists then $\{a_1, a_2, \ldots, a_n\}$ is linearly dependent and f does not lie in $\text{span}\{a_1, a_2, \ldots, a_n\}$. The system is called inconsistent. If more than one solution x exists then again $\{a_1, a_2, \ldots, a_n\}$ is linearly dependent but now $f \in \text{span}\{a_1, a_2, \ldots, a_n\}$. Finally, a unique solution implies $\{a_1, a_2, \ldots, a_n\}$ is a basis for $\mathbb{R}^n$. Theorem 2.1.30 is stated without proof (see ?).

Theorem 2.1.30 *Given A the following are equivalent*

(i) *$Ax = f$ has a unique solution for all f.*

(ii) *$Ax = 0$ has only the trivial solution $x = 0$.*

(iii) *A^{-1} exists.*

(iv) $det(A) \neq 0$.

(v) The eigenvalues of A *are nonzero.*

Problems

2.1.1 Find AB and BA if

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ -1 & -2 & -3 & -4 \\ 5 & -6 & -7 & -8 \end{bmatrix}.$$

Conclude that in general $AB \neq BA$.

2.1.2 Show that the inverse of

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 6 \end{bmatrix} \text{ is } A^{-1} = \begin{bmatrix} -2 & 0 & 1 \\ 0 & 3 & -2 \\ 1 & -2 & 1 \end{bmatrix}.$$

2.1.3 Find the eigenvalues and eigenvectors of the matrix A given in Problem 2.1.2

2.1.4 Show that the matrix B in Definition 2.1.19 is unique. Hint: assume that B is a right inverse, i.e. $AB = I$, and C is a left inverse, i.e. $CA = I$. Show that $B = C$.

2.1.5 Show that $d(A)$ is the product of the eigenvalues of A. Hint: evaluate $p(0)$ where $p(\lambda) = d(A - \lambda I)$.

2.1.6 If λ is an eigenvalue of A with eigenvector x show that λ^k is an eigenvalue of A^k with eigenvector x.

2.1.7 If λ is an eigenvalue of the nonsingular matrix A then $1/\lambda$ is an eigenvalue of A^{-1}. Prove.

2.1.8 Show that the spectrums of A^T and A are identical.

2.1.9 Prove that the eigenvalues of a triangular matrix A_n are its diagonal entries, hence,

$$d(A_n) = \prod_{i=1}^{n} a_{i,i} \ .$$

2.1.10 Prove that if any diagonal element of

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ 0 & a_{2,2} & a_{2,3} \\ 0 & 0 & a_{3,3} \end{bmatrix}$$

is zero, then the columns of A are linearly dependent.

2.1.11 Let $D(\alpha; i)$ denote the matrix obtained from I by replacing the 1 in the (i,i) position of I by α. Show that the operation $D(\alpha; i)A$ replaces row i of A by α row i while the operation $AD(\alpha; i)$ replaces column i of A by α column i.

2.1.12 Let $E(\alpha; i, j)$ denote the matrix obtained from I by replacing the zero in the (i, j)-th position of I by α. Show that the operation $E(\alpha; i, j)A$ replaces row i of A by row $i + \alpha$ row j; the operation $AE(\alpha; i, j)$ replaces column j of A by column $j + \alpha$ column i.

2.2 Forward and Backward Substitution

The solution of $Ax = f$ is greatly simplified when A is either upper or lower triangular. Let us illustrate with an example,

$$\begin{bmatrix} l_{1,1} & 0 & 0 \\ l_{2,1} & l_{2,2} & 0 \\ l_{3,1} & l_{3,2} & l_{3,3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}, l_{i,i} \neq 0 \ i = 1, 2, 3. \qquad (2.2.1)$$

The unkown x_1 is determined from the first equation

$$l_{1,1}x_1 = a_1, \text{ or } x_1 = a_1/l_{1,1}.$$

The second equation is

$$l_{2,1}x_1 + l_{2,2}x_2 = a_2, \text{ or } x_2 = (a_2 - l_{2,1}x_1)/l_{2,2}$$

which is quickly evaluated since x_1 is already known. Similarly

$$l_{3,1}x_1 + l_{3,2}x_2 + l_{3,3}x_3 = a_3, \text{ or } x_3 = (a_3 - l_{3,1}x_1 - l_{3,2}x_2)/l_{3,3}.$$

In general, if $L = [l_{i,j}]$ is an $n \times n$ lower triangular matrix the system $Lx = a$ is solved via forward substitution.

Algorithm 2.2.1 Forward Substitution

$$\begin{array}{l} \text{if } l_{1,1} = 0 \rightarrow \text{exit} \\ x_1 = a_1/l_{1,1} \\ i = 2, \ldots, n; 1 \\ \quad \text{if } l_{i,i} = 0 \rightarrow \text{exit} \\ \quad x_i = (a_i - \sum_{j=1}^{i-1} l_{i,j}x_j)/l_{i,i} \end{array}$$

Forward substitution is straightforward and results in a unique solution as long as no $l_{i,i}$ vanishes. If at least one of the $l_{i,i}$ is zero the method fails.

We next consider solving $Uy = b$, $U = [u_{i,j}]$, $u_{i,i} \neq 0$, $i = 1, 2, \ldots, n$. Thus we again examine a simple system of order 3.

$$\begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} \\ 0 & u_{2,2} & u_{2,3} \\ 0 & 0 & u_{3,3} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

We first solve for y_3, then for y_2, and finally for y_1:

$$\begin{array}{l} y_3 = b_3/u_{3,3} \\ y_2 = (b_2 - u_{2,3}y_3)/u_{2,2} \\ y_1 = (b_1 - u_{1,2}y_2 - u_{1,3}y_3)/u_{1,1} \end{array}$$

Again problems arise if for some i, $u_{i,i} = 0$. The procedure shown here is a specific example of backward substitution. Problem 2.2.1 asks the student to sketch out the algorithm for an $n \times n$ upper triangular matrix.

Problems

2.2.1 Produce an algorithm for backward substitution.

2.2.2 Consider $f(x) = \int x^n e^{ax}\, dx$. Repeated integrations by parts establish that $f(x) = p(x)e^{ax}$ where $p(x) = c_0 + c_1 x + \cdots + c_n x^n$. Since we know that

$$e^{-ax} f'(x) = x^n$$

it is possible to determine $c_0, c_1, \ldots, c_n$ by equating powers of x. This leads to a system in the unknowns $c_0, c_1, \ldots, c_n$. Develop its general form. For $a = 2, n = 4$ solve the particular linear system.

2.3 Gaussian Elimination and the *LU* Method

An elementary course in linear algebra usually offers Gaussian elimination and Cramer's rule as the two options for solving $Ax = f$. Numerically Cramer's rule is so costly that its use is restricted to a theoretical tool. Versions of Gausssian elimination, on the other hand, remain as applicable in today's computer environment as in the time of Gauss. The object of elimination algorithms is to reduce the coefficient matrix to triangular form. Minor extensions of the algorithms factor A, or a simply related matrix, into the product of lower and upper triangular matrices.

2.3.1 Gaussian Elimination

As the next few pages illustrate, the elimination procedure answers theoretical consistency and uniqueness questions assuming calculations are performed in exact arithmetic. Computationally this is insufficient; however, before addressing the added problems that occur due to machine implementation those underlying the algorithm take precedence.

When solving $Ax = f$ with Gaussian elimination f is augmented to A creating an $n \times (n+1)$ matrix $[A \vdots f]$. Suppose, for example, that the system is

$$\begin{array}{ccccccc} 2x & & & - & 3z & = & 3 \\ 4x & + & y & - & 5z & = & 7 \\ 2x & - & 3y & - & z & = & 5 \end{array}$$

then

$$[A \vdots f] = \begin{bmatrix} 2 & 0 & -3 & \vdots & 3 \\ 4 & 1 & -5 & \vdots & 7 \\ 2 & -3 & -1 & \vdots & 5 \end{bmatrix}.$$

Subtracting appropriate multiples of row one from rows two and three reduces $[A \vdots f]$ to

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

The number in the $(1,1)$ position is called the first pivot. Performing one sweep of Gaussian elimination refers to the process of eliminating the elements below the pivot by adding multiples of the i-th row to rows $i+1, \ldots, n$. With respect to the above example, the sweep around the $(2,2)$ position is accomplished by adding 3 times row 2 to row 3. This yields the configuration

$$\begin{bmatrix} 2 & 0 & -3 & \vdots & 3 \\ 0 & 1 & 1 & \vdots & 1 \\ 0 & 0 & 5 & \vdots & 5 \end{bmatrix}.$$

Backward substitution gives $z = 1, y = 0$, and $x = 3$.

The method above may fail but in such a manner that only a minor revision is needed. As an example, the augmented matrix corresponding to

$$\begin{array}{rcrcrcr} x & + & y & + & z & = & 2 \\ 3x & + & 3y & + & z & = & 0 \\ 5x & + & 3y & + & z & = & 2 \end{array}$$

reduces to

$$\begin{bmatrix} 1 & 1 & 1 & \vdots & 2 \\ 3 & 3 & 1 & \vdots & 0 \\ 5 & 3 & 1 & \vdots & 2 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & 1 & \vdots & 2 \\ 0 & 0 & -2 & \vdots & -6 \\ 0 & -2 & -4 & \vdots & -8 \end{bmatrix}$$

after one pivot. The zero in the $(2,2)$ position raises a momentary worry but fortunately the order of the equations is theoretically irrelevant so that interchanging the second and third equations furnishes a quick fix. That is,

$$\begin{bmatrix} 1 & 1 & 1 & \vdots & 2 \\ 0 & 0 & -2 & \vdots & -6 \\ 0 & -2 & -4 & \vdots & -8 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & 1 & \vdots & 2 \\ 0 & -2 & -4 & \vdots & -8 \\ 0 & 0 & -2 & \vdots & -6 \end{bmatrix},$$

which yields $z = 3$, $y = -2$, and $x = 1$.

While the last difficulty admits a simple bookkeeping cure, other problems are without remedy under Gaussian elimination. Consider the system

$$\begin{array}{rcrcrcr} x & + & y & + & z & = & 6 \\ 2x & + & 2y & + & z & = & 11 \\ 4x & + & 4y & + & 3z & = & * \end{array}$$

where for a moment * is an unspecified constant. Proceeding with the elimination routine,

$$\left[\begin{array}{ccc:r} 1 & 1 & 1 & 6 \\ 2 & 2 & 1 & 11 \\ 4 & 4 & 3 & * \end{array}\right] \to \left[\begin{array}{ccr:r} 1 & 1 & 1 & 6 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & -1 & ** \end{array}\right].$$

Notice that unlike the previous example, exchanging rows two and three does not introduce a nonzero pivot in the $(2, 2)$ position. At this stage most programmed algorithms flag an error that A is singular and the routine has failed. By hand it is possible to go the additional step

$$\left[\begin{array}{ccr:r} 1 & 1 & 1 & 6 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & -1 & ** \end{array}\right] \left[\begin{array}{ccr:r} 1 & 1 & 1 & 6 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & *** \end{array}\right].$$

Two cases occur. In the first, $* = 23$, $*** = 0$, and the system is consistent with solutions $z = 1$, $x + y = 5$. Setting $y = t$ the solutions expressed parametrically are

$$x = 5 - t, y = t, z = 1$$

which is the equation of a line in $\mathbb{R}^3$. If $***$ is nonzero then the system is inconsistent. We may interpret this in two ways: either $[6, 11, *]^T \notin \text{span}\{[1, 2, 4]^T, [1, 1, 3]^T\}$ or the three planes represented by the equations do not intersect in a common point. Unfortunately, a programmed elimination routine cannot perform this analysis and hand computation for large systems is not feasible. Singular value decomposition is an option discussed in Section 2.10.

2.3.2 The LU Method

For the remainder of this section A is assumed nonsingular. The LU method may be described via matrices as follows. In order to solve

$$Ax = f \tag{2.3.1}$$

we first set

$$A = LU \tag{2.3.2}$$

where L and U are lower and upper triangular matrices, respectively. Substituting (2.3.2) into (2.3.1) gives

$$LUx = f. \tag{2.3.3}$$

Now setting $y = Ux$ we solve the lower triangular system

$$Ly = f \tag{2.3.4}$$

with forward substitution then follow up by solving

$$Ux = y \tag{2.3.5}$$

with backward substitution. If neither $d(L)$ nor $d(U)$ vanishes then (2.3.4) and (2.3.5) have unique solutions.

The LU method's chief asset is readily seen when solutions for several right hand sides are sought. Given that the LU decomposition is known, finding x for each f uses $O(n^2)$ operations (see Problem 2.3.4) versus the $O(n^3)$ operations associated with Gaussian elimination. Of course, the start up cost of factoring A must be considered. Further, as yet an unanswered concern is whether or not A admits a factorization.

The answer to this last question is elegantly simple when A reduces to upper triangular form without row interchanges. In this case we may explain elimination via a series of elementary matrix operations and in so doing immediately derive the LU decomposition. The form of those operations is given by the following.

Definition 2.3.1 *The operation $E(\alpha; k, l)$ replaces row l of the $n \times n$ matrix A by row $k + \alpha$ row l. As an $n \times n$ matrix $E(\alpha; k, l)$ is given by*

$$E(\alpha; k, l) = \begin{cases} 1 & i = j \\ \alpha & i = k, j = l \\ 0 & \text{otherwise} \end{cases} .$$

Let us suppose that A has undergone $p-1$ steps of Gaussian elimination so that its configuration is

$$A_{p-1} = \begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \cdots & \cdots & \cdots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & \cdots & \cdots & \cdots & a_{2,n}^{(2)} \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & 0 & a_{p,p}^{(p)} & \cdots & a_{p,n}^{(p)} \\ \vdots & & \vdots & a_{p+1,p}^{(p)} & \cdots & a_{p+1,n}^{(p)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{n,p}^{(p)} & \cdots & a_{n,n}^{(p)} \end{bmatrix}$$

where the notation $a_{i,j}^{(k)}$ stands for the element replacing $a_{i,j}$ after $k-1$ elimination steps. Since we are assuming $a_{p,p}^{(p)} \neq 0$ we can define the multipliers

$$m_{i,p} = \frac{a_{i,p}^{(p)}}{a_{p,p}^{(p)}}\ , i = p+1, \ldots, n.$$

Notice that operating on A_{p-1} with $E(-m_{p+1,p}; p+1, p)$ reduces the element in the $(p+1, p)$ position to zero; that is,

$$E(-m_{p+1,p}; p+1, p)A_{p-1} = \begin{bmatrix} a_{1,1}^{(1)} & \cdots & \cdots & \cdots & \cdots & a_{1,n}^{(1)} \\ & \ddots & & & & \vdots \\ & & a_{p,p}^{(p)} & a_{p,p+1}^{(p)} & \cdots & a_{p,n}^{(p)} \\ & & 0 & a_{p+1,p+1}^{(p+1)} & \cdots & a_{p+1,n}^{(p+1)} \\ & & a_{p+2,p}^{(p)} & a_{p+2,p+1}^{(p)} & \cdots & a_{p+2,n}^{(p)} \\ & & \vdots & & & \vdots \\ & & a_{n,p}^{(p)} & \cdots & \cdots & a_{n,n}^{(p)} \end{bmatrix}.$$

Applying this concept to the remaining rows $p+2, \ldots, n$, the p-th elimination step is defined by

$$A_p = E_p A_{p-1}$$

where

$$\begin{aligned} E_p &= E(-m_{n,p}; n, p) \cdots E(-m_{p+1,p}; p+1, p) \\ &= \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & -m_{p+1,p} & 1 & & & \\ & & -m_{p+2,p} & & \ddots & & \\ & & \vdots & & & \ddots & \\ & & -m_{n,p} & & & & 1 \end{bmatrix}. \end{aligned} \tag{2.3.6}$$

Example 2.3.2 Suppose that

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ -3 & -17 & 1 & 2 \\ 4 & -17 & 8 & -5 \\ 0 & -5 & -2 & 1 \end{bmatrix}$$

then

$$\begin{aligned} A_3 &= E_3 E_2 E_1 A \\ &= E_3 E_2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 \\ -4 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 \\ -3 & -17 & 1 & 2 \\ 4 & -17 & 8 & -5 \\ 0 & -5 & 2 & 1 \end{bmatrix} \\ &= E_3 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -3/2 & 1 & 0 \\ 0 & -5/14 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & -14 & 4 & 2 \\ 0 & -21 & 4 & -5 \\ 0 & -5 & -2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -12/7 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & -14 & 4 & 2 \\ 0 & 0 & -2 & -8 \\ 0 & 0 & -24/7 & 2/7 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & -14 & 4 & 2 \\ 0 & 0 & -2 & -8 \\ 0 & 0 & 0 & 14 \end{bmatrix}. \end{aligned}$$

Clearly the matrices E_p, $p = 1, 2, \ldots, n-1$, retain pivoting information.

Theorem 2.3.3 *If Gaussian elimination requires no row exchanges then the nonsingular matrix A reduces to upper triangular form via $E_{n-1} \cdots E_2 E_1 A = U$ where the diagonal elements of U are precisely the pivots.*

We next turn our attention to constructing a lower triangular matrix L such that $A = LU$. A direct consequence of Theorem 2.3.3 is

$$\begin{aligned} A &= (E_{n-1}E_{n-2}\cdots E_1)^{-1}U \\ &= E_1^{-1}E_2^{-1}\cdots E_{n-1}^{-1}U \end{aligned}$$

Theorem 2.3.4 *If E_p is as in (2.3.6) then*

$$E_p^{-1} = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & m_{p+1,p} & 1 & & \\ & & m_{p+2,p} & & \ddots & \\ & & \vdots & & & \ddots \\ & & m_{n,p} & & & & 1 \end{bmatrix}.$$

Theorem 2.3.5 *The product $E_1^{-1}E_2^{-1}\cdots E_{n-1}^{-1}$ is the lower triangular matrix L where*

$$L = \begin{bmatrix} 1 & & & & & \\ m_{2,1} & 1 & & & & \\ m_{3,1} & m_{3,2} & 1 & & & \\ \vdots & \vdots & \ddots & \ddots & & \\ \vdots & \vdots & & \ddots & \ddots & \\ m_{n,1} & m_{n,2} & \cdots & \cdots & m_{n,n-1} & 1 \end{bmatrix}.$$

Notice that (2.3.2) may be regarded as a system of n^2 equations with $n^2 + n$ unknowns; however, if L must satisfy $l_{i,i} = 1, i = 1, 2, \ldots, n$, then the number of unkowns is reduced to n^2. Hence $A = (E_{n-1}\cdots E_2E_1)^{-1}U = LU$ is a unique factorization under the conditions stated throughout.

As a practical matter we may store both L and U by overwriting A. Thus when no row interchanges are necessary the final array stored in A is

$$A = \begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \cdots & \cdots & a_{1,n}^{(1)} \\ m_{2,1} & a_{2,2}^{(2)} & \cdots & \cdots & a_{2,n}^{(2)} \\ m_{3,1} & m_{3,2} & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ m_{n,1} & \cdots & \cdots & m_{n,n-1} & a_{n,n}^{(n-1)} \end{bmatrix}.$$

The upper triangular portion of A is U while L corresponds to the lower portion of A with 1's substituted on the main diagonal.

Example 2.3.6 For A as in Example 2.3.1??,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 \\ 4 & 3/2 & 1 & 0 \\ 0 & 5/14 & 12/7 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & -14 & 4 & 2 \\ 0 & 0 & -2 & -8 \\ 0 & 0 & 0 & 14 \end{bmatrix}$$

which would be stored as

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ -3 & -14 & 4 & 2 \\ 4 & 3/2 & -2 & -8 \\ 0 & 5/14 & 12/7 & 14 \end{bmatrix}.$$

Problems

2.3.1 If A is $n \times n$ write out an algorithm to solve $Ax = f$ using Gaussian elimination. Assume that row interchanges may be necessary.

2.3.2 Determine the number of multiplications/divisions and additions/subtractions needed to solve $Ax = f$ using Gaussian elimination.

2.3.3 An $n \times n$ matrix A is called tridiagonal if $a_{i,j} = 0$ whenever $|i-j| > 1$. The form of A is

$$\begin{bmatrix} a_1 & b_1 & & & & \\ c_2 & a_2 & b_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & & c_n & a_n \end{bmatrix}.$$

Modify your Gaussian elimination algorithm of Problem 2.3.1 to solve $Ax = f$. Your algorithm should take advantage of storage savings. What is the algorithms operation count?

2.3.4 Prove Theorem 2.3.4 .

2.3.5 Prove Theorem 2.3.5 .

2.3.6 Given that A is of order n and that $A = LU$ is known, calculate the number of multiplications/divisions and additions/subtractions needed to solve $Ax = f$.

2.3.7 Find the value of * for which

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & -1 \\ 3 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \\ * \end{bmatrix}$$

has a solution. Illustrate the solution geometrically.

2.3.8 For

$$A = \begin{bmatrix} \omega & 3 & 1 \\ 2 & -1 & 2\omega \\ 1 & 4 & \omega \end{bmatrix}, \ f = \begin{bmatrix} 5 \\ 3 \\ 6 \end{bmatrix}$$

the system $Ax = f$ has a unique solution when $\omega \neq \pm 1$, no solution when $\omega = -1$, and infinitely many solutions when $\omega = 1$. Show this.

2.3.9 Factor A into LU and solve the system

$$Ax = \begin{bmatrix} 2 & 3 & 3 \\ 0 & 5 & 7 \\ 6 & 9 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 5 \end{bmatrix}.$$

2.4 The *LU* Method with Row Interchanges

In the last section we saw that when no row exhanges are needed solving $Ax = f$ with the LU method is equivalent to Gaussian elimination. However, if A is nonsingular but a zero pivot occurs then at first glance factorization appears to fail and we must resort to elimination. In this section we adapt the LU method to permit row interchanges, not only for the purpose of circumventing zero pivots but also, as experience has shown, as a means of combatting round off error. A major difference between the theoretical versus numerical solution of a linear system is that the ordering of the equations and, indeed, even the ordering of the unknowns is significant.

Example 2.4.1 If

$$A = \begin{bmatrix} 1.10 & 0.50 & 1.00 \\ 2.40 & 1.10 & 3.00 \\ -1.00 & 0.00 & 1.00 \end{bmatrix}$$

then without row interchanges

$$L = \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 2.18 & 1.00 & 0.00 \\ -0.91 & 50.00 & 1.00 \end{bmatrix}$$

and

$$U = \begin{bmatrix} 1.10 & 0.50 & 1.00 \\ 0.00 & 0.01 & 0.82 \\ 0.00 & 0.00 & -39.00 \end{bmatrix}.$$

Solving $Ax = b$ where $b = [-.65, -3.12, -2.60]^T$ yields $\hat{x} = [.41, 2.00, -2.10]^T$ while the true solution is $x = [0.50, 1.80, -2.10]^T$. Exchanging equations two and three alters the coefficient matrix to

$$\begin{aligned} B &= \begin{bmatrix} 1.10 & 0.50 & 1.00 \\ -1.00 & 0.00 & 1.00 \\ 2.40 & 1.10 & 3.00 \end{bmatrix} \\ &= \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ -0.91 & 1.00 & 0.00 \\ 2.18 & 0.02 & 1.00 \end{bmatrix} \begin{bmatrix} 1.10 & 0.50 & 1.00 \\ 0.00 & 0.45 & 1.91 \\ 0.00 & 0.00 & 0.78 \end{bmatrix}. \end{aligned}$$

Now solving $Bx = c$ where $c = [-.65, -2.60, -3.12]^T$, $\hat{x} = [.49, 1.82, -2.10]^T$.

The previous example illustrates that factoring A into LU and then solving $LUx = b$ is less accurate than starting with $Bx = c$. Notice that the nonzero off diagonal entries of U are of the same order of magnitude but the diagonal elements vary by four orders of magnitude. This discrepancy also shows up in the nonzero elements of L making the situation ripe for round off errors. On the other hand the LU decomposition of B indicates almost no variation in orders of magnitude. Pivoting strategies, that is the selective interchange of rows and columns, seek to minimize magnitude differences.

Glancing at the matrix A is Example 2.4.1 there is no particular reason to suspect problems beforehand. Only during elimination does the difficulty present itself. It is only practical to pivot when the need appears at each elimination step. Therefore we need a means to keep track of pivoting. For instance the matrix B in Example 2.4.1 is simply related to A by

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} A.$$

The matrix relating the two is a permutation matrix.

Definition 2.4.2 *A permutation matrix is a square matrix P with exactly one unit in each row and column, the remaining elements being zero. The* *notation*

$$P = \begin{bmatrix} 1 & 2 & 3 & \cdots & n \\ j_1 & j_2 & j_3 & \cdots & j_n \end{bmatrix} \tag{2.4.1}$$

is used where $(j_1, j_2, \ldots, j_n)$ *is a permutation of the integers* $(1, 2, \ldots, n)$, *and where each* (i, j_i) *element of* P *is* 1, *the remaining positions being zero.*

Example 2.4.3 The matrix

$$P = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

is a permutation matrix of order 4. In the notation of (2.4.1)

$$P = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{bmatrix}.$$

The operation PA of the permutation matrix P on a square matrix A interchanges rows of A while the operation AP interchanges columns of A. For example, if A is 4 x 4 and P is the matrix in Example 2.4.3 then PA simultaneously transforms

$$\begin{array}{lcl} \text{row } 1 & \longrightarrow & \text{row } 3 \\ \text{row } 2 & \longrightarrow & \text{row } 1 \\ \text{row } 3 & \longrightarrow & \text{row } 4 \\ \text{row } 4 & \longrightarrow & \text{row } 2. \end{array}$$

Similarly AP takes

$$\begin{array}{lcl} \text{column } 1 & \longrightarrow & \text{column } 3 \\ \text{column } 2 & \longrightarrow & \text{column } 1 \\ \text{column } 3 & \longrightarrow & \text{column } 4 \\ \text{column } 4 & \longrightarrow & \text{column } 2. \end{array}$$

Curiously, for the same P the operations P^TA and AP^T, respectively, interchange

$$\begin{array}{lcl} \text{row } 1 & \longrightarrow & \text{row } 2 \\ \text{row } 2 & \longrightarrow & \text{row } 4 \\ \text{row } 3 & \longrightarrow & \text{row } 1 \\ \text{row } 4 & \longrightarrow & \text{row } 3 \end{array}$$

and

$$\begin{array}{lcl} \text{column } 1 & \longrightarrow & \text{column } 2 \\ \text{column } 2 & \longrightarrow & \text{column } 4 \\ \text{column } 3 & \longrightarrow & \text{column } 1 \\ \text{column } 4 & \longrightarrow & \text{column } 3 \end{array}$$

hence P^T reverses the operation P. Indeed, it is trivial to verify that for a square matrix A, $A = P^T PA = PP^T A = APP^T = AP^T P$.

Theorem 2.4.4 *Let P be a permutation matrix. Then*

$$P^T P = PP^T = I. \tag{2.4.2}$$

The proof is left as an exercise.

Decomposition methods that allow for row interchanges rest on the following.

Theorem 2.4.5 *If A is nonsingular then there exists a permutation matrix P such that*

$$PA = LU. \tag{2.4.3}$$

Ignoring for a moment the actual construction of P, L and U, let us rewrite $Ax = f$ as

$$PAx = LUx = Pf. \tag{2.4.4}$$

Once P is known (2.4.4) admits immediate solution via forward and backward substitution.

Our goal is to determine a P such that stable numerical calculations ensue from the LU decomposition of PA. As a side benefit zero pivots are managed naturally in the process. An easily implemented, well accepted pivoting strategy is scaled column pivoting.

To describe the algorithm we begin by assuming the $n \times n$ matrix A has undergone $k - 1$ elimination steps and that the multipliers are stored in the portion of A that has been zeroed out. Therefore, the present configuration of A is

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & \cdots & \cdots & a_{1,n} \\ m_{2,1} & a_{2,2} & & & & \vdots \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & m_{k,k-1} & a_{k,k} & \cdots & a_{k,n} \\ \vdots & & \vdots & a_{k+1,k} & \cdots & a_{k+1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ m_{n,1} & \cdots & m_{n,k-1} & a_{n,k} & \cdots & a_{n,n} \end{bmatrix}$$

where it is understood that elements of A are overwritten as we proceed. For $i = k, k+1, \ldots, n$ let

$$d_i^{(k)} = \max_{k \le j \le n} |a_{i,j}|.$$

Scaled column pivoting selects the pivot of maximum magnitude with respect to the element of maximum absolute value in its row, in other words the first value of $i = k, k+1, \ldots, n$, for which

$$\frac{|a_{i,k}|}{d_i^{(k)}}$$

is largest. Denote this row by i_k. The next step of the algorithm exchanges rows k and i_k of A, including those columns containing the multipliers. Now continue as before.

The final configuration of A is the LU decomposition of a matrix simply related to A by a permutation matrix. As in the last section

$$U = \begin{cases} a_{i,j} & i \le j \\ 0 & \text{otherwise} \end{cases} \tag{2.4.5}$$

and

$$L = \begin{cases} a_{i,j} & i > j \\ 1 & i = j \\ 0 & \text{otherwise} \end{cases} . \tag{2.4.6}$$

It remains to establish P.

Definition 2.4.6 *Let $P(i,j)$ denote the permutation matrix obtained from I by interchanging the i-th and j-th rows of I.*

Theorem 2.4.7 *The operation $P(i,j)A$ interchanges rows i and j of A. The operation $AP(i,j)$ interchanges columns i and j of A.*

As may be expected, a general permutation matrix P is related to the operations $P(i,j)$.

Theorem 2.4.8 *Every permutation matrix P can be expressed as a finite product of elementary operations $P(i,j)$.*

Returning to our algorithm, denote $P(k, i_k)$ as the permutation matrix that exchanges rows k and $i_k \geq k$ prior to carrying out the kth elimination step on A.

Theorem 2.4.9 *If $P = P(n-1, i_{n-1})P(n-2, i_{n-2}) \cdots P(1, i_1)$ then $PA = LU$ where L and U are given by (2.4.6) and (2.4.5), respectively.*

Example 2.4.10 Suppose

$$A = \begin{bmatrix} 2 & -2 & 4 \\ 4 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Then under scaled column pivoting we exchange rows 1 and 2 and proceed with the first elimination step. At its completion the matrix stored in A is

$$\begin{bmatrix} 4 & 1 & 1 \\ 1/2 & -5/2 & 7/2 \\ 1/4 & -1/4 & -1/4 \end{bmatrix}.$$

Comparing 5/7 and 1/1, scaled column pivoting once again dictates row exchanges, this time rows 2 and 3. The final array is

$$\begin{bmatrix} 4 & 1 & 1 \\ 1/4 & -1/4 & -1/4 \\ 1/2 & 10 & 6 \end{bmatrix}.$$

Therefore

$$PA = LU$$

where

$$\begin{aligned} P &= P(2,3)P(1,2) \\ &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \\ L &= \begin{bmatrix} 1 & 0 & 0 \\ 1/4 & 1 & 0 \\ 1/2 & 10 & 1 \end{bmatrix}, \end{aligned}$$

and

$$U = \begin{bmatrix} 4 & 1 & 1 \\ 0 & -1/4 & 1/4 \\ 0 & 0 & 6 \end{bmatrix} .$$

To solve

$$Ax = \begin{bmatrix} -2 \\ 6 \\ 1 \end{bmatrix} ,$$

let

$$g = P(2,3)P(1,2) \begin{bmatrix} -2 \\ 6 \\ 1 \end{bmatrix} = P(2,3) \begin{bmatrix} 6 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 1 \\ -2 \end{bmatrix}$$

so that

$$LUx = g .$$

Now apply forward and backward substitution

A second, more costly, pivoting strategy is maximal pivoting. Before the k-th elimination step select an element

$$d^{(k)} = \max_{k \le i,j \le n} |a_{i,j}|$$

then using row and column interchanges move the element to the (k,k) position. Proceed as usual. This strategy should be reserved for systems the user suspects are particularly susceptible to round off error, such as those whose elements differ by several orders of magnitude.

For general linear systems it is virtually impossible to state a priori how much more accurate a numerical solution is when computed with a pivoting routine. There are, however, conditions under which it is known that pivoting does not significantly improve numerical approximations.

Definition 2.4.11 *The $n \times n$ matrix A is said to be strictly row diagonally dominant if for all $i = 1, 2, \ldots, n$*

$$|a_{i,i}| > \sum_{j=1, j \neq i}^{n} |a_{i,j}|.$$

Definition 2.4.12 *The $n \times n$ matrix A is strictly column diagonally dominant if for all $j = 1, 2, \ldots, n$*

$$|a_{j,j}| > \sum_{i=1, i \neq j}^{n} |a_{i,j}|.$$

Theorem 2.4.13 *If the $n \times n$ matrix A is either strictly row or column diagonally dominant then A is nonsingular. Moreover an LU decomposition of A based on Gaussian elimination can be carried out without row exchanges and is stable with respect to the growth of rounding errors.*

Definition 2.4.14 *The $n \times n$ symmetric matrix A is called positive definite if for all $x \in \mathbb{R}^n$ such that $x \neq 0$ $x^T Ax > 0$.*

Theorem 2.4.15 *If the $n \times n$ matrix A is positive definite then its eigenvalues are real and positive. Moreover A admits an LU decomposition via Gaussian elimination without row interchanges and this decomposition is stable with respect to rounding errors.*

Problems

2.4.1 Prove Theorem 2.4.4.

2.4.2 Prove Theorem 2.4.7.

2.4.3 Prove that the eigenvalues of a positive definite matrix are positive.

2.4.4 Find the LU decomposition of

$$A = \begin{bmatrix} 1 & 4 & 2 \\ -2 & -8 & 3 \\ 0 & 1 & 1 \end{bmatrix}$$

using scaled column pivoting. Solve the system $Ax = f$ where $f = [-2, 32, 1]^T$.

2.4.5 Show that the $n \times n$ positive definite matrix A satisfies the following:

(i) $a_{ii} > 0$ for $i = 1, 2, \ldots, n$

(ii) $\max_{1 \leq i \leq n} |a_{i,i}| \geq \max_{1 \leq i,j \leq n} |a_{i,j}|$

(iii) $(a_{i,j})^2 < a_{i,i} a_{j,j}$ for $i, j = 1, 2, \ldots, n$; $i \neq j$

2.4.6 Show that

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

is positive definite.

2.5 Vector and Matrix Norms

Solving $Ax = b$ in finite precision arithmetic yields a computed solution $\hat{x}$ that must be regarded as an approximate to the true solution x. We can further cloud the picture by noting that round off may not permit exact representation of A and b; hence, most often we are solving a perturbed version of the true system. Clearly we need tools to measure the error $e = x - \hat{x}$ before we can approach the problem of ensuring that the error is "small". In this section we introduce the concept of a norm in $\mathbb{R}^n$ as a measure of vector length.

Definition 2.5.1 *A norm in $\mathbb{R}^n$ is a function, $\|\cdot\|$, from $\mathbb{R}^n$ to the nonnegative real numbers such that for all $x,y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$*

(i) $\|x\| \geq 0$ *and* $\|x\| = 0 \Leftrightarrow x = 0$

(ii) $\|\alpha x\| = |\alpha|\|x\|$

(iii) $\|x + y\| \leq \|x\| + \|y\|$.

Criterion (i) implies that the length of a nonzero vector in $\mathbb{R}^n$ is always positive, (ii) guarantees that the vectors x and $-x$ have the same length, and (iii) is the familiar triangle inequality. It follows that a natural measure of the distance between two vectors x and y is $\|x - y\|$.

Although an infinity of norms exist we confine our discussion to the ℓ_1 (taxicab), ℓ_2 (Euclidean), and ℓ_∞ (max) norms.

Definition 2.5.2 *For $x = [x_1, x_2, \ldots, x_n]^T$ the ℓ_1, ℓ_2, and ℓ_∞ norms are denoted by $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ where*

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \ ,$$

$$\|x\|_2 = \left\{ \sum_{i=1}^{n} x_i^2 \right\}^{1/2} ,$$

and

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i| \ .$$

The proofs that $\|\cdot\|_1$ and $\|\cdot\|_\infty$ satisfy the definition of a norm is left as an exercise. Since the proof for $\|\cdot\|_2$ differs somewhat we include it here.

Theorem 2.5.3 *The function $\|\cdot\|_2$ is a vector norm in $\mathbb{R}^n$.*

Proof: Clearly (i) is satisfied while (ii) follows from

$$\|\alpha x\|_2 = \left\{\sum_{i=1}^{n}(\alpha x_i)^2\right\}^{1/2} = \left\{\alpha^2\sum_{i=1}^{n}x_i^2\right\}^{1/2} = |\alpha|\left\{\sum_{i=1}^{n}x_i^2\right\}^{1/2} = |\alpha|\|x\|_2 \,.$$

To establish (iii) we rely on the Cauchy-Schwarz inequality

$$\sum_{i=1}^{n}|x_iy_i| \leq \left\{\sum_{i=1}^{n}x_i^2\right\}^{1/2}\left\{\sum_{i=1}^{n}y_i^2\right\}^{1/2} \tag{2.5.1}$$

that the reader is asked to verify in Problem 2.5.3. Proceeding under the assumption that the Cauchy-Schwarz inequality holds we have

$$\begin{aligned}\|x+y\|_x^2 &= \sum_{i=1}^{n}x_i^2 + 2\sum_{i=1}^{n}x_iy_i + \sum_{i=1}^{n}y_i^2 \\ &\leq \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2 \\ &\leq (\|x\|_2 + \|y\|_2)^2 \,.\end{aligned}$$

Taking square roots concludes the proof.

Example 2.5.4 For $x = [2, -5, 3/2]^T$, $\|x\|_1 = 17/2$, $\|x\|_2 = 5\sqrt{5}/2$, and $\|x\|_\infty = 5$.

At this juncture we pause to offer geometric interpretations of the ℓ_1 and ℓ_2 unit speres in $\mathbb{R}^2$, that is the values $x \in \mathbb{R}^2$ such that $\|x\|_p = 1$ for $p = 1, 2$. If $x = [x, y]^T$ then the Euclidean unit sphere is $\{(x, y) : x^2 + y^2 = 1\}$, a circle of radius 1 centered at the origin (see Figure 2.5 below). The ℓ_1 sphere follows almost as rapidly in we note that

$$\begin{aligned}\{(x,y) : |x| + |y| = 1\} &= \{(x,y) : \text{ for } x \geq 0, y \geq 0, x + y = 1\} \\ &\quad \cup \{(x,y) : \text{ for } x > 0, y < 0, x - y = 1\} \\ &\quad \cup \{(x,y) : \text{ for } x < 0, y > 0, -x + y = 1\} \\ &\quad \cup \{(x,y) : \text{ for } x < 0, y < 0, x + y = -1\} \,.\end{aligned}$$

Figure 2.5 illustrates the diamond shaped ℓ_1 sphere.

Example 2.5.4 and Figure 2.5 clearly show that barring a few exceptional vectors the choice of norm leads to different values of $\|x\|$. If we are trying to control $\|e\|$ where e is some error this variation between norms could be disturbing however the succeeding definition and theorem imply that in essence we can pick a norm to suit our convenience.

Figure 2.1: The ℓ_2 and ℓ_1 unit spheres

Definition 2.5.5 *The norms* $\|\cdot\|$ *and* $\|\cdot\|'$ *are called equivalent if there exist constants* α *and* β *such that for all* $x \in \mathbb{R}^n$

$$\alpha\|x\| \leq \|x\|' \leq \beta\|x\| \ .$$

Equivalence of norms is reflexive in the sense that the statement above implies there exist constants γ and θ such that

$$\gamma\|x\|' \leq \|x\| \leq \theta\|x\|' \ .$$

Theorem 2.5.6 *The* ℓ_1, ℓ_2, *and* ℓ_∞ *norms are pairwise equivalent.*

Proof: See ?.

Having defined equivalent norms it is now convenient to mention convergent sequences, a central concept for the algorithms included in the next two sections.

Definition 2.5.7 *The sequence* $\left\{x^{(k)}\right\}_{k=1}^{\infty}$ *in* $\mathbb{R}^n$ *is said to converge to* x *in norm* $\|\cdot\|$ *if for all* $\varepsilon > 0$ *there exists* $N(\varepsilon)$ *such that*

$$\|x^{(k)} - x\| < \varepsilon \text{ for all } k > N(\varepsilon).$$

Theorem 2.5.8 *If the vector sequence* $x^{(k)} \to x$ *in any one of the norms* ℓ_1, ℓ_2, *or* ℓ_∞ *then it converges to the identical limit in the remaining norms.*

Proof: By equivalence of norms.

Example 2.5.9 The sequence $x^{(k)} = [e^{2-k}, -2/k^2, 1/k]^T$ converges to $x = [0,0,0]^T$.

Besides measuring the length of a vector in $\mathbb{R}^n$ we need some measure of size for an $n \times n$ matrix. At the outset a matrix norm apears very similar to a vector norm with one added constraint.

Definition 2.5.10 *Let A and B be real $n \times n$ matrices and $\alpha \in \mathbb{R}$. The real valued function $\|\cdot\|$ is said to be a matrix norm if $\|\cdot\|$ satisfies the following*

(i) $\|A\| \geq 0$ *and* $\|A\| = 0 \Leftrightarrow A = 0$

(ii) $\|\alpha A\| = |\alpha| \|A\|$

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

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

One standard measuring $\|A\|$ is by comparison of the size of Ax with x:

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} . \tag{2.5.2}$$

This is called an induced or natural norm. An inequality immediately follows for induced norms

$$\|Ax\| \leq \|A\| \|x\| .$$

It can be shown that (2.5.2) is equivalent to

$$\|A\| = \max_{\|x\|=1} \|Ax\| . \tag{2.5.3}$$

Fortunately, calculating $\|A\|_p$ is fairly straightforward when $p = 1, \infty$.

Theorem 2.5.11 $\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{i,j}|$.

Proof: The proof follows in two parts. First, let $x \in \mathbb{R}^n$ be such that $\|x\|_1 = 1$. Then

$$\begin{aligned} \|Ax\|_1 &= \sum_{i=1}^{n} |(Ax)_i| \\ &= \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{i,j} x_j| \end{aligned}$$

$$\leq \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{i,j}|$$
$$\leq \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{i,j}|$$

hence

$$\max_{\|x\|_1=1} \|Ax\|_1 \leq \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{i,j}| \, .$$

To establish the inequality in the other direction let m correspond to a column such that $\sum_{i=1}^{n} |a_{i,j}|$ is a maximum. Define the particular vector $y = [y_1, y_2, \ldots, y_n]^T$ by

$$y_j = \begin{cases} 0 & j \neq m \\ 1 & j = m \end{cases} \, .$$

Then

$$\begin{aligned} \max_{\|x\|_1=1} \|Ax\|_1 &\geq \|Ay\|_1 \\ &= \sum_{i=1}^{n} |a_{i,m}| \\ &= \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{i,j}| \end{aligned}$$

which completes the proof.

Theorem 2.5.12 $\|A\|_\infty = \max_{1\leq i\leq n} \sum_{i=1}^{n} |a_{i,j}|$.

Proof: See Problem 2.5.6.

The ℓ_2 norm of a matrix A is seldom computed via (2.5.3) simply because it is too difficult. However, there is an intimate connection between the singular values of A and its ℓ_2 norm.

Definition 2.5.13 *The singular values of a real $n \times n$ matrix A are the eigenvalues of $A^T A$.*

Example 2.5.14 For

$$A = \begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix}$$

$$A^TA = \begin{bmatrix} 10 & -10 \\ -10 & 20 \end{bmatrix} .$$

Therefore, the singular values of A are the roots of $p(\lambda) = \lambda^2 - 30\lambda + 100$, or $\lambda = 15 \pm 5\sqrt{5}$.

Definition 2.5.15 *The spectral radius* $\rho(A)$ *of a matrix* A *is given by*

$$\rho(A) = \left\{ \max_{\lambda} |\lambda| \ : \ Ax = \lambda x, x \neq 0 \right\} .$$

Theorem 2.5.16 *If* A *is a real* $n \times n$ *matrix then*

(i) $\|A\|_2 = \sqrt{\rho(A^TA)}$.

(ii) $\rho(A) \leq \|A\|$ *for any induced norm* $\|\cdot\|$.

Proof: For a proof of (i) see ?.

(ii) Let λ be an eigenvalue of A with unit eigenvector x. Then

$$|\lambda| = \|\lambda x\| \leq \|A\|$$

hence

$$\rho(A) \leq \|A\| .$$

Problems

2.5.1 Show that $\|\cdot\|_1$ satisfies the definition of a norm.

2.5.2 Show that $\|\cdot\|_\infty$ satisfies the definition of a norm.

2.5.3 Prove the Cauchy-Schwarz inequality (see 2.5.1). Hint: evaluate $\|x - \lambda y\|^2$ and select a useful value of λ.

2.5.4 Find $\|x\|_1$, $\|x\|_2$, and $\|x\|_\infty$ for $x = [2\sin\theta, 2\cos\theta, 8]^T$.

2.5.5 Find $\lim_{k\to\infty} x^{(k)}$ where $x^{(k)} = [e^{1/k}, \sqrt{k^2+k} - k, (\sin k)/k]^T$.

2.5.6 Prove Theorem 2.5.12 .

2.5.7 Calculate $\|A\|_1$, $\|A\|_2$, and $\|A\|_\infty$ for

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

2.5.8 If λ is an eigenvalue of the nonsingular matrix A show that for any natural norm $\|\cdot\|$, $(1/\|A^{-1}\|) \leq |\lambda|$.

2.5.9 Show that if A is symmetric then $\|A\|_2 = \max\{|\lambda|; Ax = \lambda x\}$.

2.6 Residual Refinement

A dilemma inherent in the numerical solution of linear systems is that the error $e = x - \hat{x}$ can only be determined if the true solution x is known. A sensible task is to estimate either $\|e\|$ or $\|e\|/\|x\|$. Most often the residual vector $r = b - A\hat{x}$ serves as a reliable indicator of e in the sense that $\|r\|/\|b\|$ and $\|e\|/\|x\|$ are within a few orders of magnitude of each other. Intuitively it seems quite reasonable that if the computed solution $\hat{x}$ closely satisfies the equation $Ax = b$ then $\hat{x}$ should closely approximate x. Unfortunately there are systems for which our intuition is at fault as the next example illustrates.

Example 2.6.1 Let

$$A = \begin{bmatrix} 1.001 & 1.999 \\ .999 & 2.001 \end{bmatrix}, \; b_1 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \quad \text{and } b_2 = \begin{bmatrix} -.3 \\ .3 \end{bmatrix}.$$

First we consider the system $Ax = b_1$ and note that its true solution is $x = [1,1]^T$. If $\hat{x} = [1.001, .999]^T$ then $e = [.001, -.001]^T$ and $r = [-.001, -.001]^T$. Measuring with the ℓ_1 norm, $\|r\|/\|b\| = .0003$ and $\|e\|/\|x\| = .0005$. Now if for some bizarre reason $\hat{x} = [3,0]^T$ then $r = [.003, -.003]^T$, $e = [2,-1]^T$ and $\|r\|/\|b\| = .001$ while $\|e\|/\|x\| = .333$. Clearly a small relative residual does not imply a small relative error. Likewise, if we turn our attention to the system $Ax = b_2$, a large relative residual may occur with a small relative error. Here the true solution is $x = [-200, 100]^T$. If $\hat{x} = [-202, 99]^T$ then $\|r\|/\|b\| = 13.33$ while $\|e\|/\|x\| = .01$.

Depending on lady luck to produce better results than expected is unwise, hence, we need some means of deciding when it is safe to rely on the residual. For a system of order 2 it is possible to foresee problems graphically. The figure below shows two lines that are nearly parallel. For a goodly range of values (x_1, x_2) the lines remain close even though the point of intersection may be some distance away.

Figure 2.2: Nearly Parallel Lines

Potential difficulties with large systems are detected via norms. We begin with the fundamental relationship between residual and error

$$r = b - A\hat{x} = Ax - A\hat{x} = Ae \tag{2.6.1}$$

or equivalently

$$e = A^{-1}r \tag{2.6.2}$$

where we are assuming A^{-1} exists. Applying norm properties to (2.6.2) yields

$$\begin{aligned} \|e\| &\leq \|A^{-1}\|\|r\| \\ &= \|A^{-1}\|\|Ax\|\frac{\|r\|}{\|b\|} \\ &\leq \|A^{-1}\|\|A\|\|x\|\frac{\|r\|}{\|b\|} \end{aligned}$$

or,in terms of relative error,

$$\frac{\|e\|}{\|x\|} \leq \|A^{-1}\|\|A\|\frac{\|r\|}{\|b\|}\,.$$

Similarly we can bound the relative error from below as follows:

$$\|A\|\|e\| \geq \|Ae\| \geq \|r\|$$

which implies

$$\begin{aligned}\frac{\|e\|}{\|x\|} &\geq \frac{1}{\|x\|\|A\|}\|r\| \\ &= \frac{1}{\|A^{-1}b\|\|A\|}\|r\| \\ &\geq \frac{1}{\|A^{-1}\|\|A\|}\frac{\|r\|}{\|b\|} .\end{aligned}$$

Definition 2.6.2 *The condition number of a matrix A is given by*

$$\mathit{cond}(A) = \|A^{-1}\|\|A\| .$$

Combining the results above, we have

$$\frac{1}{\mathrm{cond}(A)}\frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|r\|}{\|b\|} . \qquad (2.6.3)$$

By Problem 2.6.1, $\mathrm{cond}(A) \geq 1$. When $\mathrm{cond}(A)$ is near 1 the relative error is constrained to closely follow the relative residual while a large value of $\mathrm{cond}(A)$, especially with respect to machine precision, allows for wide variation between relative error and residual.

Example 2.6.3 (Example 2.6.1 revisited) For

$$A = \begin{bmatrix} 1.001 & 1.999 \\ .999 & 2.001 \end{bmatrix}$$

the ℓ_1 norm, $\|A\|_1 = 4$, is small; however,

$$A^{-1} = \begin{bmatrix} 333.6 & -333.2 \\ -166.5 & 166.9 \end{bmatrix}$$

has a much larger ℓ_1 norm, $\|A^{-1}\|_1 = 500.1$, which combines to give $\mathrm{cond}_1(A) = 2000$. For the right hand side b_1 and $\hat{x} = [1.001, .999]^T$ the relative error and residual are the same order of magnitude. When $\hat{x} = [3, 0]^T$ the relative error is roughly 300 times larger than the relative residual. Finally for b_2 and $\hat{x} = [-202, 99]^T$, $\|e\|/\|x\|$ is about 1/1000-th of $\|r\|/\|b\|$. The variation is clearly attributable to $\mathrm{cond}_1(A)$.

Two items we wish to mention are first that the value of $\mathrm{cond}(A)$ depends on the norm used and, second, that in practice $\mathrm{cond}(A)$ is at best estimated since computing A^{-1} is costly. Even if we attempt to calculate A^{-1} it may be unreliable for precisely the same reason solutions of $Ax = b$ are. Estimates of cond(A) are given by **IMSL**, **MATLAB**, and **EISPACK** routines. We elaborate more shortly.

Iterative refinement is a technique applied to overcome numerical errors due to a poorly conditioned coefficient matrix. Consider the linear system

$$Ay = r \ . \tag{2.6.4}$$

If $r = 0$ then $y = 0$ and $\hat{x}$ is without error. If $r \neq 0$ then $y \neq 0$ and

$$y = A^{-1}r = x - \hat{x} \ .$$

Of course, we do not actually calculate y but rather $\hat{y}$. Nevertheless, $\hat{x} + \hat{y}$ should be an improved approximation to x.

Example 2.6.4 Consider solving the system

$$\begin{bmatrix} 9 & -36 & 30 \\ -36 & 192 & -180 \\ 30 & -180 & 180 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -15 \\ 96 \\ -90 \end{bmatrix}$$

with two decimal rounding arithmetic. The LU decomposition of the coefficient matrix,

$$\begin{bmatrix} 9 & -36 & 30 \\ -36 & 192 & -180 \\ 30 & -180 & 180 \end{bmatrix} = \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ -4.00 & 1.00 & 0.00 \\ 3.33 & -1.25 & 1.00 \end{bmatrix} \begin{bmatrix} 9.00 & -36.00 & 30.00 \\ 0.00 & 48.00 & -60.00 \\ 0.00 & 0.00 & 5.10 \end{bmatrix}$$

leads to the approximate $\hat{x} = [2.94, 1.96, .97]^T$. The residual is $r = b - A\hat{x} = [0, .12, 0]^T$. Using the LU decomposition again, we solve $Ay = r$ and get $\hat{y} = [.06, .04, .03]^T$. Therefore, an updated approximate is $\hat{\hat{x}} = [3.00, 2.00, 1.00]^T$. We recognize this vector as the true solution.

When one step of iterative refinement is insufficient to produce a satisfactory result we merely define the iterative process

$$\hat{x}^{(k)} = \hat{x}^{(k-1)} + \hat{y}^{(k-1)}$$

where $\hat{y}^{(k-1)}$ is the numerical solution of

$$Ay = r^{(k-1)}$$

$$r^{(k-1)} = b - A\hat{x}^{(k-1)} \ .$$

In practice $\hat{y}^{(k)}$ is calculated using double precision arithmetic because its entries become successively smaller as the iteration continues. Some stopping criteria are necessary such as a limit on the maximum number of iterations or

$$\|\hat{x}^{(k)} - \hat{x}^{(k-1)}\| \leq \varepsilon$$

where ε is a user supplied tolerance.

Residual refinement tends to be costly and, as a result, is used only when the coefficient matrix is suspected of being ill-conditioned. The cost does not justify the expense for well-behaved matrices. This is true even if we perform an LU decomposition of A so that the solution of (2.6.4) is an $O(n^2)$ versus an $O(n^3)$ operational step.

We close with an estimate of $\text{cond}(A)$. If t-digit arithmetic is used to perform a Gaussian type LU decomposition for A then ii can be shown [ref]

$$\text{cond}(A) \approx \frac{\|\hat{y}\|}{\|\hat{x}\|} 10^t \ .$$

Problems

2.6.1 Show $\text{cond}(A) \geq 1$ for any natural norm.

2.6.2 Show that if B is singular then

$$\frac{1}{\text{cond}(A)} \leq \frac{\|A - B\|}{\|A\|}$$

2.6.3 The condition number also measures the sensitivity of A to round off error. Consider the system

$$\begin{bmatrix} 1 & 3 \\ 1.0001 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4.0001 \end{bmatrix}$$

whose solution is $x = [1, 1]^T$. If A is perturbed slightly so that we are solving

$$\begin{bmatrix} 1 & 3 \\ .9999 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} 4 \\ 4.0001 \end{bmatrix}$$

what is the solution now (use four digit rounding)? Comment on $\text{cond}(A)$. Estimate $\text{cond}(A)$ using the result of Problem 2.6.2.

2.6.4 Solve the system

$$\begin{array}{rcrcrcrcl} 6.99x_1 & + & 10.01x_2 & + & 6.99x_3 & + & 8.01x_4 & = & 32 \\ 5.01x_1 & + & 6.99x_2 & + & 5.01x_3 & + & 5.99x_4 & = & 23 \\ 5.99x_1 & + & 8.01x_2 & + & 8.99x_3 & + & 10.01x_4 & = & 33 \\ 5.01x_1 & + & 6.99x_2 & + & 10.01x_3 & + & 8.99x_4 & = & 31 \end{array}$$

with an LU decomposition truncating (not rounding) all calculations to two digits. Compare your results with the true solution $x_1 = x_2 = x_3 = x_4 = 1$. Perform one iteration of residual refinement. Estimate $\text{cond}(A)$.

2.7 Iterative Methods

2.7.1 Motivation

It is not uncommon for a physical problem to give rise to extremely large sparse linear systems, say on an order of 10^3 to 10^6. For systems of this order, direct methods based on Gaussian elimination may require too much machine time and storage to be feasible. Further, a direct method may fill in many of the previously zero entries thereby creating more sources of round off error. Iterative methods are a viable alternative if the cost of each iteration is significantly less than elimination and if the sequence generated converges very rapidly.

Our first concern is motivating iterative methods and, hopefully, in the process indicating circumstances that guarantee the methods convergence.

Many linear systems arising in applications are naturally ordered so that the coefficient of x_k in the k-th equation is large in magnitude relative to the other coefficients in that equation. If this is true then rewriting

$$Ax = b$$

as

$$\begin{aligned} x_1 &= \frac{1}{a_{1,1}}(b_1 - a_{1,2}x_2 - a_{1,3}x_3 - \cdots - a_{1,n}x_n) \\ x_2 &= \frac{1}{a_{2,2}}(b_2 - a_{2,1}x_1 - a_{2,3}x_3 - \cdots - a_{2,n}x_n) \\ &\vdots \\ x_n &= \frac{1}{a_{n,n}}(b_n - a_{n,1}x_1 - a_{n,2}x_2 - \cdots - a_{n,n-1}x_{n-1}) \end{aligned}$$

a good initial guess for x_i, $i = 1, 2, \ldots, n$, is

$$x_i^{(0)} = \frac{1}{a_{i,i}} b_i \ .$$

In vector form

$$x^{(0)} = D^{-1}b$$

where D is a diagonal matrix whose diagonal elements are those of A.

Next we must determine how to update an approximate $x^{(k-1)}$ to $x^{(k)}$. Residual refinement would dictate

$$x^{(k)} = x^{(k-1)} + y^{(k-1)}$$

where $Ay^{(k-1)} = r^{(k-1)}$. To avoid solving $Ay^{(k-1)} = r^{(k-1)}$ we approximate $y^{(k-1)}$ by $D^{-1}r^{(k-1)}$ which yields the iterative equation

$$x^{(k)} = x^{(k-1)} + D^{-1}r^{(k-1)}$$

or componentwise

$$x_i^{(k)} = x_i^{(k-1)} + \frac{1}{a_{i,i}} r_i^{(k-1)} \,. \tag{2.7.1}$$

There is some leeway in defining $r_i^{(k-1)}$. For instance, the most obvious choice is

$$r_i^{(k-1)} = b_i - \sum_{i=1}^{n} a_{i,j} x_j^{(k-1)}$$

which substituted into 2.7.1 gives

$$\begin{aligned} x_i^{(k)} &= x_i^{(k-1)} + \frac{1}{a_{i,i}} \left(b_i - \sum_{i=1}^{n} a_{i,j} x_j^{(k-1)} \right) && (2.7.2) \\ &= -\frac{1}{a_{i,i}} \sum_{j=1, j\neq i}^{n} a_{i,j} x_j^{(k-1)} + \frac{1}{a_{i,i}} b_i \,. && (2.7.3) \end{aligned}$$

If we let

$$L = \begin{bmatrix} 0 & & & & \\ -a_{2,1} & 0 & & & \\ -a_{3,1} & -a_{3,2} & 0 & & \\ \vdots & & \ddots & \ddots & \\ -a_{n,1} & -a_{n,2} & \cdots & -a_{n,n-1} & 0 \end{bmatrix}$$

and

$$U = \begin{bmatrix} 0 & -a_{1,2} & -a_{1,3} & \cdots & -a_{1,n} \\ & 0 & -a_{2,3} & \cdots & -a_{2,n} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & -a_{n-1,n} \\ & & & & 0 \end{bmatrix}$$

then the matrix equation represented by 2.7.3 is

$$x^{(k)} = T_J x^{(k-1)} + c$$

where

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

and

$$c = D^{-1}b \; .$$

This is called the Jacobi iterative method.

A second choice for residual utilizes the known components of $x_i^{(k)}$ via

$$r_i^{(k-1)} = b_i - \sum_{j=1}^{i-1} a_{i,j} x_j^{(k)} - \sum_{j=i}^{n} a_{i,j} x_j^{(k-1)} \; . \qquad (2.7.4)$$

Again we substitute into 2.7.1 and after simplifying find

$$x_i^{(k)} = -\frac{1}{a_{i,i}} \left(\sum_{j=1}^{i-1} a_{i,j} x_j^{(k)} + \sum_{j=i+1}^{n} a_{i,j} x_j^{(k-1)} \right) + \frac{1}{a_{i,i}} b_i$$

or

$$\sum_{j=1}^{i} a_{i,j} x_j^{(k)} = - \sum_{j=i+1}^{n} a_{i,j} x_j^{(k-1)} + b_i \; .$$

Letting $i = 1, 2, \ldots, n$ we have the matrix equation

$$(-L + D)x^{(k)} = Ux^{(k-1)} + b$$

or

$$x^{(k)} = T_G x^{(k-1)} + c$$

where

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

and

$$c = (D - L)^{-1} b \; .$$

The method outlined above is referred to as Gauss-Seidel iteration. While there is no general rule stating that Gauss-Seidel iteration converges faster than Jacobi iteration this is usually true.

A third class of iterative techniques are called relaxation methods. One such method uses the same form of the residual as Gauss-Seidel but the iterative equation,

$$x_i^{(k)} = x_i^{(k-1)} + \frac{\omega}{a_{i,i}} r_i^{(k-1)} \; , \qquad (2.7.5)$$

now allows for a parameter ω whose purpose is to accelerate convergence. Rearranging terms in 2.7.5 gives

$$a_{i,i} x_i^{(k)} + \omega \sum_{j=1}^{i-1} a_{i,j} x_j^{(k)} = (1-\omega) a_{i,i} x_i^{(k-1)} - \omega \sum_{j=i+1}^{n} a_{i,j} x_i^{(k-1)} + \omega b_i$$

or, running through $i = 1, 2, \ldots, n$

$$x^{(k)} = T_\omega x^{(k-1)} + c$$

where

$$T_\omega = (D - \omega L)^{-1}((1-\omega)D + \omega U)$$

and

$$c = \omega(D - \omega L)^{-1} b \; .$$

If $\omega = 1$ the relaxation method is simply Gauss-Seidel. When $\omega < 1$ we are said to be under-relaxing, while $1 < \omega$ corresponds to over-relaxation. For $0 < \omega < 2$ we refer to the method as SOR for successive over-relaxation.

2.7.2 Convergence of Iterative Methods

While it may not be possible to establish a priori how fast an iterative method converges there are circumstances that at least guarantee convergence.

Definition 2.7.1 *An $n \times n$ matrix A is convergent if*

$$\lim_{k\to\infty} (A^k)_{i,j} = 0 \text{ for } 1 \le i, j \le n$$

where $(A^k)_{i,j}$ refers to the (i,j)-th element of A^k.

Theorem 2.7.2 *For an $n \times n$ matrix A the following are equivalent*

(i) *A is convergent*

(ii) $\lim_{n\to\infty} \|A^n\| = 0$ *for some natural norm* $\|\cdot\|$

(iii) $\rho(A) < 1$.

Lemma 2.7.3 *If for the $n \times n$ matrix A, $\rho(A) < 1$ then $(I - A)^{-1}$ exists and admits the convergent expansion*

$$(I - A)^{-1} = I + A + A^2 + \cdots$$

Theorem 2.7.4 *For any initial guess $x^{(0)}$ the sequence given by*

$$x^{(k)} = T x^{(k-1)} + c \; , k = 1, 2, \ldots$$

where $c \neq 0$ converges to the unique solution of $x = Tx + c$ if and only if $\rho(T) < 1$.

Proof: Stepping back through the first $k-1$ iterates we have

$$x^{(k)} = T^k x^{(0)} + (T^{k-1} + \cdots + T + I)c\,. \tag{2.7.6}$$

If $\rho(T) < 1$ then taking the limit as $k \to \infty$ of 2.7.6 and applying the lemma gives

$$\begin{aligned} \lim_{k\to\infty} x^{(k)} &= \lim_{k\to\infty} T^k x^{(0)} + \lim_{k\to\infty} (T^{k-1} + \cdots + T + I)c \\ &= 0 + (I-T)^{-1}c \\ &= x\,. \end{aligned}$$

The proof of the converse is left as an exercise.

Not only does the spectral radius determine whether or not an iterative method converges it also provides a good estimate of how rapid convergence is.

Corollary 2.7.5 *Suppose that* $\|\cdot\|$ *is a natural norm. If* $\|T\| < 1$ *the sequence* $x^{(k)} = Tx^{(k-1)} + c$, $c \neq 0$, *converges to* x *for any initial guess* $x^{(0)}$. *Moreover,* $\|x - x^{(k)}\|$ *is bounded by*

$$\|x - x^{(k)}\| \le \|T\|^k \|x - x^{(0)}\| \tag{2.7.7}$$

or

$$\|x - x^{(k)}\| \le \frac{\|T\|^k}{1 - \|T\|} \|x^{(1)} - x^{(0)}\|\,. \tag{2.7.8}$$

Proof: By definition

$$\begin{aligned} x - x^{(k)} &= Tx - Tx^{(k-1)} \\ &= T\left(x - x^{(k)}\right) + T\left(x^{(k)} - x^{(k-1)}\right) \end{aligned}$$

which implies

$$\begin{aligned} (I-T)(x - x^{(k)}) &= T\left(x^{(k)} - x^{(k-1)}\right) \\ &= T^k\left(x^{(1)} - x^{(0)}\right)\,. \end{aligned}$$

Taking norms

$$\begin{aligned} \|x - x^{(k)}\| &\le \frac{\|T\|^k}{\|I - T\|} \|x^{(1)} - x^{(0)}\| \\ &\le \frac{\|T\|^k}{1 - \|T\|} \|x^{(1)} - x^{(0)}\|\,. \end{aligned}$$

If we replace $\|T\|$ in Corollary 2.7.5 with $\rho(T)$ then

$$\|x - x^{(k)}\| \approx \rho(T)^k \|x - x^{(0)}\| \; .$$

Therefore, when $\rho(T)$ is close to 1 we expect slow convergence. In particular, we can estimate the number of iterations needed to obtain t digits of relative error by

$$\rho(T)^k \leq 10^{-t}$$

or, rather,

$$k \geq \frac{-t}{\log_{10} \rho(T)} \; .$$

The object of relaxation is to select the parameter ω so that the spectral radius of $T_\omega = (D - \omega L)^{-1}((1-\omega)D + \omega U)$ is a minimum. It is known [see?] that when $0 < \omega < 1$ it is possible to force convergence of systems for which Gauss-Seidel does not converge. When $1 < \omega < 2$ over-relaxations attempts to accelerate the convergence of Gauss-Seidel iteration.

Guaranteeing convergence of iterative methods based on properties of the iterative matrix T is all good and fine but sufficient conditions for A also exist. These tend to be easier to verify but are fairly restrictive.

Theorem 2.7.6 *If A is strictly diagonally dominant then for any initial guess $x^{(0)}$ the sequences defined by the Jacobi and Guass-Seidel iterative methods converge to the true solution of $Ax = b$.*

Before stating the analgous theorem for the SOR method the next theorem addresses the restriction $0 < \omega < 2$.

Theorem 2.7.7 *If the diagonal elements of A are nonzero then $\rho(T_\omega) \geq |\omega - 1|$;hence, a necessary condition of $\rho(T_\omega) < 1$ is $0 < \omega < 2$.*

Theorem 2.7.8 *If A is positive definite and $0 < \omega < 2$ then the SOR method converges to the true solution of $Ax = b$ for any initial guess $x^{(0)}$.*

Problems

2.7.1 Show that

$$A = \begin{bmatrix} 1 & 0 \\ \frac{1}{9} & \frac{1}{3} \end{bmatrix}$$

is not convergent but

$$B = \begin{bmatrix} \frac{1}{3} & 0 \\ 9 & \frac{1}{3} \end{bmatrix}$$

is convergent. In general, for a lower triangular matrix what is a necessary condition to guarantee convergence?

2.7.2 Show that if A is strictly diagonally dominant then $\|T_J\|_\infty < 1$.

2.7.3 Show that A is a convergent matrix if and only if $\lim_{k\to\infty} A^k x \to 0$ for every vector x.

2.7.4 Consider the system

$$\begin{bmatrix} 1.4 & 2.3 & 3.7 \\ 3.3 & 1.6 & 4.3 \\ 2.5 & 1.9 & 4.1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6.5 \\ 10.3 \\ 8.8 \end{bmatrix} .$$

Comment on applying Gauss-Seidel iteration to the system. Experiment with SOR by varying the parameter ω.

Chapter 3

The Solution of Nonlinear Equations

Introduction - Sketching $y = f(x)$

The most rudimentary method of sketching $y = f(x)$ consists of selecting a few points x_i in the domain of f, evaluating $y_i = f(x_i)$, plotting the coordinates (x_i, y_i), and then tracing a curve through these points. In regions where f changes rapidly a heavier concentration of points (x_i, y_i) is necessary to capture the behavior of the function.

A more sophisticated approach to curve tracing utilizes the following concepts.

a. The x intercepts of f; that is, the values ξ such that $f(\xi) = 0$.

b. Vertical asumptotes; that is, lines $x = a$, $|a| < \infty$, such that f is unbounded as $x \to a$.

c. Asymptotes of infinity; that is, curves that describe the behavior of f as x becomes unbounded.

d. Critical values of f; that is, values η in the domain of f where either $f'(\eta) = 0$ or $f'(\eta)$ fails to exist. In the former case the tangent line at $(\eta, f(\eta))$ is horizontal while in the latter the tangent line is vertical. If in addition the sign of $f'(x)$ changes on open intervals to the left and right of η then the point $(\eta, f(\eta))$ corresponds to a relative maximum or minimum.

e. Inflection points of f; that is, points $(\zeta, f(\zeta))$ such that either $f''(\zeta) = 0$ or $f''(\zeta)$ does not exist and, further, $f''(x)$ changes sign on open intervals to the left and right of ζ.

Example 3.0.1 Consider $y = f(x)$ where

$$\begin{aligned} f(x) &= \frac{1}{2}x^2 - e^{-x} \\ f'(x) &= x + e^{-x} \\ f''(x) &= 1 - e^{-x}. \end{aligned}$$

We first note that $f(0)f(1) < 0$ hence by the Intermediate Value Theorem f has a root $\xi \in (0,1)$. Since $f'(x) > 0$ for all x (why?) this value ξ is the unique root of f and, moreover, there are no relative maxima or minima. From $f''(x)$ we glean the information that $(0,-1)$ is an inflection point. Finally, note that the behavior of f as $x \to \infty$ is dominated by the term $\frac{1}{2}x^2$ while as $x \to -\infty$ f closely follows $-e^{-x}$. Combining this information yields the rough sketch of f shown in Figure 3.0.1 below. If we were now to numerically sketch the graph we would concentrate points near $x = 0$ to detect the change in concavity and concentrate points on the interval $(0,1)$ to locate the root ξ.

Figure 3.0.1: The graph $f(x) = \frac{1}{2}x^2 - e^{-x}$.

This example called on two concepts. The first is the notion of solving $f(x) = 0$ (and possibly $f'(x) = 0$, $f''(x) = 0$). Finding roots of functions is the subject of the present chapter. The second notion is that of asymptotics. In all but relatively trivial examples we defer this topic to Chapter 6.

Problems

3.0.1 Sketch the graph of each of the following equations.

a. $y = x(x^2 - 1)$ (This has 3 real roots.)

b. $1000y = 1000x(x^2 - 1) - 89$ (This has 3 real roots, 2 close together.)

c. $1000y = 1000x(x^2 - 1) - 90$ (This has one real root.)

3.0.2 Sketch the graph of the equation

$$xy^2 - 2xy + y - 1 = 0$$

and use it to show that the equation

$$x^7 - 2x^4 + x^3 - 1 = 0$$

only has one real root.

3.0.3 Descartes rule of signs states that a polynomial with real coefficients can only have as many positive real roots as there are changes of signs in the coefficients, although it may have $2k$ fewer roots than the number of changes of sign, where $k \geq 0$ is an integer. How many real roots can the polynomial

$$p(x) = x^7 - c^5x^2 + c^7$$

have if (a) $c > 0$, (b) $c < 0$?

3.1 Elementary One-Dimensional Methods

Finding a satisfactory value of ξ such that $f(\xi) = 0$ involves a miriad of considerations. For instance, a rough sketch of $y = f(x)$ may reveal that the root we seek corresponds to a change in sign of $f(x)$. If so and if only a crude approximation to the root, accurate to two or three significant digits, suffices then the simple method of bisections outlined below adequately suits the setting. If greater accuracy is sought we turn to the more sophisticated secant method or Newton's method. In exchange for greater accuracy, these methods have the disadvantage that they must be started sufficiently close to ξ before they converge.

For each of the methods we describe we sketch the algorithm (not code) for obtaining an aproximate value of ξ. In each algorithm, the user must specify the value of the positive number ε that is the desired accuracy of the computed root. In order to avoid difficulties due to round off error, ε should be larger than the magnitude of the error in the newly computed approximation in each cycle, assuming all values are exact at the start of the cycle. In other words, it is unreasonable to insist that ε be smaller than the precision of machine calculations. Finally, the user must also supply an upper limit on the number of iterations the algorithm performs.

3.1.1 The Method of Bisections

The bisection method is based on successive application of the Intermediate Value Theorem.

Example 3.1.1 Consider finding an approximate value of ξ such that

$$f(x) = x^3 - 2x - 3 = 0.$$

Either by means of the sketch in Figure 3.1.1 or by noting that $f(1)f(2) < 0$, it is clear that f must be zero for some $\xi \in (1,2)$. Since $f'(x) > 0$ on $(1,2)$, $f(x) = 0$ exactly once on $(1,2)$.

We evaluate $f(x)$ at the midpoint of $[1,2]$ to get

$$f(1.5) = -2.625 < 0 < 1 = f(2)$$

hence, the root ξ lies in the smaller interval $(1.5, 2)$. Next evaluating f at the midpoint 1.75 we have

$$f(1.75) \approx -1.14 < 0 < 1 = f(2)$$

and we now have that $1.75 < \xi < 2$; that is, the approximation $\xi = 1.875$ is in error by less than .125.

In general if f is continuous on some interval $[a,b]$ and $f(a)f(b) < 0$ then in the absence of round off error the bisection method is guaranteed to converge to within ε of a root $\xi \in (a,b)$.

Algorithm 3.1.2 Bisection Algorithm

$c = (a+b)/2$
if $b - c \leq \varepsilon$ then accept $c \to$ exit
    otherwise
    if $f(c)f(a) < 0$
        then $b = c$
        otherwise $a = c$

Figure 3.1.1: Sketch of $f(x) = x^3 - 2x - 3$.

Example 3.1.3 Referring to the function $f(x) = \frac{1}{2}x^2 - e^{-x}$ in Example 3.0.1 we tabulate the results in Table 3.1.1.

We remark that the bisection method converges even if more than one root of $f(x)$ is present in the interval. Of course, a rough sketch of $f(x)$ is usually sufficient for the user to select an interval $[a, b]$ containing a unique root.

There are disadvantages associated with the bisection method. Chief amongst them is that the algorithm converges very slowly when compared to the methods discussed next. Further, because the only information the algorithm uses is the sign of f at various points, good intermediate approximations are ignored. Not to abandon the algorithm altogether, bisection is often used to narrow down the location of a root before switching to a faster iterative technique. These faster techniques must, among other restrictions, be started sufficiently close to a root to guarantee convergence.

3.1.2 The Secant Method

In order to numerically find a root ξ of $f(x)$ faster than the bisection method more information about the values of f must be used at each step. For example, if by the Intermediate Value Theorem we know a unique root ξ of $f(x)$ exists on $[a, b] = [x_0, x_1]$ then it is possible to estimate ξ with the x-intercept of the secant line joining $(x_0, f(x_0))$ and $(x_1, f(x_1))$. See Figure

Table 3.1.1: The Bisection Algorithm applied to $f(x) = \frac{1}{2}x^2 - e^{-x}$.

n	c_n	Error Bound	$f(c_n)$
	$a = 0.5$		-.4815 3066
	$b = 1.0$		.1321 2055
1	.75	.25	-.1911 1655
2	.875	.125	-.0340 0495
3	.9375	.0625	.0478 4749
4	.9062 5	.0312 5	.0066 0800
5	.8906 25	.0156 25	-.0137 9272
6	.8984 375	.0078 125	-.0036 1045
7	.9023 4375	.0039 0625	.0014 9424
8	.9003 9062	.0019 5312	-.0010 5923
9	.9013 6718	.0009 7656	.0020 1722
10	.9008 7890	.0004 8828	-.0004 2107

3.1.2 below. Using the points $(x_0, f(x_0))$, $(x_1, f(x_1))$, and $(x_2, 0)$ we can equate two expressions for the slope of the secant line, namely,

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{0 - f(x_1)}{x_2 - x_1}.$$

Now solving for x_2 gives

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1). \tag{3.1.1}$$

One interpretation of this expression is that x_2 is x_1 plus a correction term. Another explanation follows by rewriting (3.1.1) as a weighted average between x_0 and x_1 the weights being $|f(x_0)|$ and $|f(x_1)|$, that is,

$$x_2 = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)} \tag{3.1.2}$$

Intuitively from this explanation and the graph below, in general we expect faster convergence of the secant method than the bisection method.

Algorithm 3.1.4 Secant Algorithm

Figure 3.1.2: The first iterate of the secant method.

$F = f(x_0)$
$G = f(x_1)$
for $n = 1, 2, \ldots$
$$x_{n+1} = \frac{x_n - x_{n-1}}{G - F} G$$
if $|x_{n+1} - x_n| \leq \varepsilon$ then accept $x_{n+1} \rightarrow$ exit
otherwise
$F = G$
$G = f(x_{n+1})$

Example 3.1.5 Again we have $f(x) = \frac{1}{2}x^2 - e^{-x}$, $x_0 = 0.5$, and $x_1 = 1.0$. Unlike the bisection algorithm applied to the same problem each iterate of $f(x_n)$ gains at least a decimal point.

Several remarks are in order. First we choose to iterate using the expression
$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) \tag{3.1.3}$$
rather than
$$x_{n+1} = \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} \,. \tag{3.1.4}$$
Although (3.1.3) and (3.1.4) are algebraically equivalent, in practice (3.1.4) is more prone to round off error than (3.1.3). Our second remark is reminiscent

Figure 3.1.3: The secant method.

Table 3.1.2: The secant method applied to $f(x) = \frac{1}{2}x^2 - e^{-x}$.

n	x_{n+1}	$\lvert x_{n+1} - x_n \rvert$	$f(x_{n+1})$
	$x_0 = 0.5$		
	$x_1 = 1.0$		
1	.8923 4882		-.0115 4911
2	.9010 0252	.0086 7652	-.0002 5948
3	.9012 0142	.0001 9889	.0000 0050

of the conditioning discussion in Chapter 2. That is the secant method may converge to a point such that $|f(x)|$ is small and the iterates are "close", however, this may not approximate the root accurately. The figure below attempts to illustrate the situation. Notice that Table 3.1.1 includes an error bound column while the best we can do with Table table 3.1.2 is include a column tabulating $|x_{n+1} - x_n|$. Fortunately there are circumstances that guarantee the iterates converge to the true solution ξ of $f(x) = 0$. It can be shown that when f' exists and is non-zero in a neighborhood of ξ then by selecting x_0 and x_1 sufficiently close to ξ the secant method converges to ξ.

We close our discussion of the secant method by noting that some authors refer to the algorithm as the method of false position, or *regula falsi* method, while for others the method of false position and the secant method are

Figure 3.1.4:

distinct. A third variant of the secant method is the modified *regula falsi* method. These variants are discussed in the problems at the end of the section.

3.1.3 Newton's Method

The student of calculus knows a limiting process replaces the secant line by the tangent line. Therefore a natural adaptation of the secant method is to begin with a single point x_0 and plot its tangent line. As before, using the points $(x_0, f(x_0))$ and $(x_1, 0)$ along with $f'(x_0)$ we can calculate the x-intercept of the tangent line by equating slopes. That is,

$$f'(x_0) = \frac{x_0 - x_1}{f(x_0) - 0}$$

yields

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} .$$

An alternative derivation uses the Taylor series development of $f(x)$ about $x = x_n$,

$$f(x) = f(x_n) + (x - x_n)f'(x_n) + \frac{(x - x_n)^2}{2!}f''(\eta)$$

Figure 3.1.5: The first iterate of Newton's Method.

where η is between x and x_n. If we let $x = \xi$ and use $f(\xi) = 0$ then

$$\xi = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{(\xi - x_n)^2}{2!} \frac{f''(\eta)}{f'(x_n)} .$$

Hence the last term represents error and we iterate using

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} .$$

Algorithm 3.1.6 Newton's Algorithm

for $n = 1, 2, \ldots$
  $x_n = x_{n-1} - f(x_{n-1})/f'(x_{n-1})$
  if $|x_n - x_{n-1}| < \varepsilon$ then accept $x_n \to$ exit

Example 3.1.7 For $f(x) = \frac{1}{2}x^2 - e^{-x}$ Table 3.1.3 indicates that Newton's method outperforms the already rapid secant method.

In Section 3.4 we establish that if $f''(x)$ exists in a neighborhood N of a solution ξ of $f(x) = 0$, if $f'(x) \neq 0$ in N, and if x_0 is sufficiently close to ξ then the iterates of Newton's method converge to ξ.

Table 3.1.3: Newton's method applied to $f(x) = \frac{1}{2}x^2 - e^{-x}$.

n	x_n	$\lvert x_n - x_{n-1}\rvert$	$f(x_n)$
	$x_0 = 1$		
1	.9034 1213	.0965 8786	.0028 9198
2	.9012 0214	.0022 0999	.0000 0145
3	.9012 0103	.0000 0111	-1.4×10^{-12}

We note that among the bisection, secant, and Newton's methods, Newton's method converged fastest for the example discussed. However, Newton's method requires two function evaluations per step, that of $f(x)$ and $f'(x)$, while the other two methods require only one function evaluation per iterate. Therefore, an iterate of Newton's method is more costly besides requiring explicit knowledge of $f'(x)$. Indeed, it is shown in Section 3.5 that the rate of convergence per function evaluation is fastest for the secant method. Of the three, only Newton's method is applicable for complex roots.

Problems

3.1.1 a. By sketching a graph or otherwise, find an interval containing the real positive root of the function $x^2 - 3x - 1$.

b. Compute the real positive root of $x^2 - 3x - 1 = 0$ correct to 4 decimals by the method of bisections.

3.1.2 In the method of bisections let $x_1, x_2, x_3, \ldots$ represent successive approximations.

a. Show that
$$|x_{i+1} - x_i| \leq \frac{|x_1 - x_2|}{2^{i-1}}.$$

b. Show that the number I of evaluations of f required to guarantee an approximation that is accurate to within ε is
$$I > \frac{\log\left[\frac{|x_1 - x_2|}{\varepsilon}\right]}{\log 2} + 1.$$

3.1.3 The bisection method is applicable when $f(a)f(b) < 0$. If $f(x)$ has more than one zero in the interval $[a, b]$ which zero, if any, does the method locate?

3.1.4 Given the following functions

$$f(x) = (3x-1)^{-1} \qquad g(x) = \cos 10x \qquad h(x) = \begin{cases} 1 & x \geq 0.3 \\ -1 & x < 0.3 \end{cases}$$

that change sign in $[0, 1]$

a. What point does the bisection algorithm locate?

b. Is this a true zero of the function?

3.1.5 Carry out 3 steps of the secant method for the problem $x^3 - 3 = 0$, with $x_0 = 0$,$x_1 = 1$.

3.1.6 Use Newton's method to calculate the unique root of

$$x + e^{-Bx^2} \cos x = 0$$

with $B > 0$ a parameter to be chosen. Use a variety of increasing values of B, for example, $B = 1, 5, 10, 25, 50$. Among the choices of x_0 used, choose $x_0 = 0$ and explain any anomalous behavior. Theoretically, the Newton method converges for any value of x_0 and B. Compare this with actual computations for larger values of B.

3.1.7 Apply the Secant method to the equation in Problem 3.1.6

3.2 The Fixed Point Method

From the point of view of understanding the convergence properties of iterative methods it is more convenient to study the equation

$$x = g(x) \tag{3.2.1}$$

rather than

$$f(x) = 0. \tag{3.2.2}$$

Figure 3.2.1: Fixed points of $y = g(x)$.

When we solve the latter equation we are said to find the roots of $f(x)$ while in the nomenclature of the former equation we are finding the fixed points of $g(x)$. Figure3.2.1 below illustrates a function with three fixed points.

It is natural to solve the equation (3.2.1) by starting with some initial approximate x_0 and forming the sequence $\{x_n\}$ using the recurrence relation

$$x_{n+1} = g(x_n)\ ,\ n = 0, 1, 2, \ldots\ . \tag{3.2.3}$$

Figure 3.2.2 illustrates that if $|g'(x)| < 1$ in a neighborhood N of the fixed point ξ and $x_0 \in N$ then $x_n \to \xi$ as $n \to \infty$. On the other hand Figure 3.2.3 shows that when $|g'(x)| > 1$ in general the sequence $\{x_n\}$ does not converge to ξ. We say "in general" because of course if $x_0 = \xi$ then convergence is trivial.

Example 3.2.1 Referring to the function $f(x) = \frac{1}{2}x^2 - e^{-x}$ in Example 3.0.1 one obvious way to convert the problem of finding ξ such that $f(\xi) = 0$ to a fixed point problem is to solve for x; that is,

$$g(x) = \sqrt{2e^{-x}}\ . \tag{3.2.4}$$

Selecting $x_0 = 1.0$ and iterating to a tolerance of $tol = .00005$ gives convergence after 11 iterates as shown in Table 3.2.1.

Another method of converting $f(x) = 0$ to a fixed point problem is simply to add x to both sides of the equation. For the particular $f(x)$ we are given

$$g(x) = x + \frac{1}{2}x^2 + e^{-x} \tag{3.2.5}$$

Figure 3.2.2: $|g'(x)| < 1 \Longrightarrow$ convergence.

Figure 3.2.3: $|g'(x)| > 1 \Longrightarrow$ possible divergence.

this proves unsuccessful as evidenced by Table 3.2.2 . The difficulty lies with the constraint on the magnitude of $g'(x)$ near $x = .9$.

Theorem 3.2.2 *Let g be defined on the interval $I \equiv [x_0 - a, x_0 + a]$, let*

$$|g(x) - g(y)| \leq \alpha |x - y| \tag{3.2.6}$$

for all x and y on this interval, where $0 \leq alpha < 1$, and let

$$|x_0 - g(x_0)| = |x_0 - x_1| \leq (1 - \alpha)a\,. \tag{3.2.7}$$

Then the sequence $\{x_p\}$ defined by $x_p = g(x_{p-1})$ converges to a number $\xi \in I$. Moreover, $\xi = g(\xi)$ is the unique solution in I of $x = g(x)$ in I and

$$|x_p - \xi| \leq \frac{\alpha}{1 - \alpha}\,|x_p - x_{p-1}| \leq \alpha^p a. \tag{3.2.8}$$

Table 3.2.1: The fixed point method applied to $g(x) = (2e^{-x})^{1/2}$.

n	x_{n-1}	$x_n = g(x_{n-1})$	$\lvert x_n - x_{n-1}\rvert$
1	1.000 000	.857 764	.142 236
2	.857 764	.920 988	.063 224
3	.920 988	.892 329	.028 659
4	.892 329	.905 208	.012 879
5	.905 208	.899 397	.005 810
6	.899 397	.902 014	.002 617
7	.902 014	.900 835	.001 179
8	.900 835	.901 366	.000 531
9	.901 366	.901 127	.000 239
10	.901 127	.901 235	.000 108
11	.901 235	.901 186	.000 049

Table 3.2.2: The fixed point method applied to $g(x) = x + \frac{1}{2}x^2 + e^{-x}$.

n	x_{n-1}	$x_n = g(x_{n-1})$	$\lvert x_n - x_{n-1}\rvert$
1	1.000 000	1.132 121	.132 121
2	1.132 121	1.450 620	.318 500
3	1.450 620	2.268 345	.817 724
4	2.268 345	4.737 555	2.469 210
5	4.737 555	15.951 008	11.213 453

Proof: We first show that $x_p \in I$ for all $n \geq 0$. This is clearly the case for $n = 0$ and $n = 1$. Now if $x_0, x_1, \ldots, x_k \in I$ where $k \geq 1$ then

$$\begin{aligned} |x_{k+1} - x_k| &= |g(x_k) - g(x_{k-1})| \\ &\leq \alpha |x_k - x_{k-1}| \\ &\leq \\ &\vdots \\ &\leq \alpha |x_1 - x_0| \\ &\leq \alpha^k (1-\alpha) a. \end{aligned} \tag{3.2.9}$$

Hence,

$$\begin{aligned} |x_{k+1} - x_0| &= \left| \sum_{j=0}^{k} (x_{j+1} - x_j) \right| \\ &\leq \sum_{j=0}^{k} |x_{j+1} - x_j| \\ &\leq \sum_{j=0}^{k} \alpha^j (1-\alpha) a \\ &= (1 - \alpha^{k+1}) a < a \end{aligned} \tag{3.2.10}$$

so that $x_{k+1} \in I$. Hence $x_p \in I$ for all $p \geq 0$.

Next we note by (3.2.10) that for $p \geq 1$ and $m \geq 10$

$$\begin{aligned} |a_{m+p} - x_m| &= \sum_{k=0}^{p-1} (x_{m+k+1} - x_{m+k}) \\ &\leq \sum_{k=0}^{p-1} |x_{m+k+1} - x_{m+k}| \\ &\leq \left[\sum_{k=0}^{p-1} \alpha^k \right] |x_{m+1} - x_m| \\ &\leq \frac{1-\alpha^p}{1-\alpha} \alpha^m (1-\alpha) |x_1 - x_0| \\ &\leq \alpha^m a. \end{aligned} \tag{3.2.11}$$

This estimate shows that the sequence $\{x_n\}$ is a Cauchy sequence that converges to some point $\xi \in I$. Letting $p \to \infty$ on the left hand side of (3.2.11) we get the error bounds

$$|\xi - x_m| \leq \frac{\alpha}{1-\alpha} |x_m - x_{m-1}| \leq \frac{\alpha^m}{1-\alpha} |x_1 - x_0| \, . \tag{3.2.12}$$

In order to show that $\xi = g(\xi)$, we note that

$$\begin{aligned} |\xi - g(\xi)| &\leq |\xi - x_p| + |x_p - g(\xi)| \\ &= |\xi - s_p| + |g(x_{p-1}) - g(\xi)| \\ &\leq |\xi - x_p| + \alpha |x_{p-1} - \xi| \\ &\leq \alpha^p a + \alpha^p a \, , \end{aligned} \tag{3.2.13}$$

and letting $p \to \infty$ on the right hand side, it follows that

$$|\xi - g(\xi)| = 0 \tag{3.2.14}$$

or, equivalently, that $\xi = g(\xi)$.

In order to prove uniqueness, let us assume that there exists an $\eta \in I$ such that $\eta = g(\eta)$ but $\eta \neq \xi$. Then, since $g(\xi) = \xi$,

$$0 < |\xi - \eta| = |g(\xi) - g(\eta)| < \alpha|\xi - \eta| < |\xi - \eta|\,,$$

which is a contradiction. Hence there is no $\eta \in I$ such that $\eta = g(\eta)$ and $\eta \neq \xi$. This completes the proof of Theorem 3.2.2.

The above theorem and proof may be used to describe several different numerical methods. It also gives an insight regarding the type of function g we want in order to achieve rapid convergence, how close to the solution ξ we should start the method, and how fast the method converges. We also obtain an *a priori* estimate of the error of x_n that depends only on x_{n-1} , x_n, and α.

In numerical analysis we sometimes hear the following statements regarding iterative methods such as Newton's method: "The method converges if it is started close enough to the solution." The following corollary states that this is true for the function g of Theorem 3.2.2

Theorem 3.2.3 *Let $\xi = g(\xi)$, let $|g'(\xi)| < 1$, and let $g'(x)$ be continuous in a neighborhood of ξ. If x_0 is sufficiently close to ξ, then the sequence $\{x_n\}$ defined by $x_n = g(x_{n-1})$ converges to ξ.*

Proof:

Problems

3.2.1 For each of the following equations determine an iteration function (and an interval I) so that the conditions of Theorem 3.2.3 are satisfied with respect to the smallest positive root.

a. $x^3 - x - 1 = 0$

b. $x - \tan x = 0$

c. $e^{-x} - \cos x = 0$

3.2.2 Determine the largest interval I with the property that for all $x_0 \in I$ fixed point iteration with the iteration function

$$g(x) = x(2 - ax)$$

converges when started with x_0. Is Theorem 3.2.3 satisfied for your choice of I? What numbers are possible limits of this iteration? Can you think of a good reason for using this particular iteration? Note that the interval depends on the constant a.

3.2.3 Same as Problem 3.2.2 but with

$$g(x) = \frac{x + a/x}{2}.$$

3.2.4 The equation $e^x - 4x^2 = 0$ has a root between $x = 4$ and $x = 5$. Show that we cannot find this root using fixed point iteration with the"natural" iteration function

$$x = \frac{1}{2}e^{x/2}.$$

Can you find an iteration function that correctly locates this root?

3.2.5 The equation $e^x - 4x^2 = 0$ also has a root between $x = 0$ and $x = 1$. Show that the iteration function $x = \frac{1}{2}e^{x/2}$ converges to this root if x_0 is chosen in the iterval $[0, 1]$.

3.3 Rate of Convergence of Iterative Methods

Consider finding a solution ξ of the equation $x = g(x)$ by generating a sequence $x_{p+1} = g(x_p)$, $p = 0, 1, 2, \ldots$. If g has a continuous derivative in a neighborhood of ξ then a Taylor series expansion about ξ gives

$$\begin{aligned} g(x) &= g(\xi) + (x - \xi)g'(\eta) & (3.3.1) \\ &= \xi + (x - \xi)g'(\eta) & (3.3.2) \end{aligned}$$

for some η between x and ξ. Substituting $x = x_{p-1}$ into (3.3.2) the equation becomes

$$x_p = \xi + (x_{p-1} - \xi)g'(\eta) \qquad (3.3.3)$$

or

$$|x_p - \xi| = |g'(\eta)|\,|x_{p-1} - \xi|\,. \tag{3.3.4}$$

Now for x_n sufficiently close to ξ (3.3.4) admits the approximate

$$|x_p - \xi| \approx |g'(\xi)|\,|x_{p-1} - \xi|\ ,\ p = 0, 1, \ldots \tag{3.3.5}$$

hence if $0 < |g'(\xi)| < 1$ then error decreases with each successive iterate.

More generally, at the solution ξ of the equation $x = g(x)$, the following may occur

$$g'(\xi) = g''(\xi) = \cdots = g^{(q-1)}(\xi) = 0 \tag{3.3.6}$$

$$g^{(q)}(\xi) \neq 0 \tag{3.3.7}$$

where q is a positive integer. In this case, the Taylor series expansion about ξ yields

$$g(x) = g(\xi) + \frac{(x-\xi)^q g^{(q)}(\eta)}{q!} \tag{3.3.8}$$

where again η is between x and ξ. Proceeding as above we have the estimate

$$|x_p - \xi| \approx \frac{g^{(q)}(\xi)|}{q!}\,|x_{p-1} - \xi|^q\ ,\ p = 1, 2, \ldots \tag{3.3.9}$$

Therefore the method is said to be of order q since for x_p sufficiently close to ξ x_p has roughly q times as many correct significant figures as x_{p-1}.

3.3.1 The Order of Newton's Method

Let us find the order q of Newton's method where recall the expression

$$x_{p+1} = x_p - \frac{f(x_p)}{f'(x_p)} \tag{3.3.10}$$

implies that the $g(x)$ we are considering is

$$g(x) = x - \frac{f(x)}{f'(x)}\,. \tag{3.3.11}$$

Differentiating $g(x)$

$$g'(x) = \frac{f(x)\,f''(x)}{(f'(x))^2} \tag{3.3.12}$$

we see that $g'(x)$ vanishes when $x = \xi$ assuming that $f'(x)$ as well as $f''(x)$ exist and are non-zero in a neighborhood N of ξ. Now checking the second derivative of $g(x)$,

$$\begin{aligned} g''(x) &= \frac{f(x)\, f^{(3)}(x)}{(f'(x))^2} + \frac{f''(x)}{f'(x)} - \frac{2f(x)(f''(x))^2}{(f'(x))^3} & (3.3.13) \\ &= \frac{f''(\xi)}{f'(\xi)} \neq 0 \; . & (3.3.14) \end{aligned}$$

Hence if $x_p \to \xi$ then

$$|x_p - \xi| \approx \left| \frac{f''(\xi)}{2f'(\xi)} \right| |x_{p-1} - \xi|^2 \; . \qquad (3.3.15)$$

In this setting, Newton's method is said to converge quadratically.

When $f'(\xi) = 0$ but $f'' \neq 0$ in a neighborhood N of ξ it is still possible to demonstrate the convergence of Newton's method. Using the Taylor series of $f(x)$ and $f'(x)$ we have

$$\begin{aligned} f(x) &= \frac{(x-\xi)^2}{x} f''(\xi) + \frac{(x-\xi)^3}{3!} f^{(3)}(\eta) & (3.3.16) \\ f'(x) &= (x-\xi) f''(\xi) + \frac{(x-\xi)^2}{2} f^{(3)}(\zeta) & (3.3.17) \end{aligned}$$

so that

$$g'(x) = \frac{f(x)\, f''(x)}{(f'(x))^2} \to \frac{1}{2} \text{ as } x \to \xi \, . \qquad (3.3.18)$$

Therefore referring to (3.3.9), when $f(\xi) = f'(\xi) = 0$, $f''(\xi) \neq 0$ Newton's method is apparently a first order method.

In general, the order of a method need not be an integer.

Definition 3.3.1 *The* order of an iterative method *that converges to ξ is defined to be the largest number q such that*

$$\limsup_{p\to\infty} \frac{|x_p - \xi|^{q'}}{|x_{p+1} - \xi|} = \infty \qquad (3.3.19)$$

for all $q' < q$.

3.3.2 The Order of the Secant Method

To deduce the order of the secant method we replace x_0, x_1, and x_2 in (3.1.1) by x_{p-1}, x_p, and x_{p+1} respectively. Then subtracting ξ from each side of (3.1.1) gives

$$\begin{aligned} \xi - x_{p+1} &= \xi - x_p + f(x_p)\,\frac{x_p - x_{p-1}}{f(x_p) - f(x_{p-1})} && (3.3.20)\\ &= (\xi - x_p)\,\frac{f[x_{p-1}, x_p] - f[x_{p-1}, \xi]}{f[x_{p-1}, x_p]} && (3.3.21)\end{aligned}$$

where we use the notation

$$f[a, b] = \frac{f(a) - f(b)}{a - b} \tag{3.3.22}$$

from Chapter ?. The right hand side of (3.3.22) further simplifies to the form

$$\xi - x_{p+1} = -(\xi - x_p)(\xi - x_{p-1})\,\frac{f[x_{p-1}, x_p, \xi]}{f[x_{p-1}, x_p]} \tag{3.3.23}$$

where

$$f[x_{p-1}, x_p, \xi] = \frac{f[x_{p-1}, x_p] - f[x_p, \xi]}{x_{p-1} - \xi}\,. \tag{3.3.24}$$

If we anticipate some results of Chapter ? and assume that $f''(x)$ exists in a neighborhood N of ξ then (3.3.21) yields

$$\xi - x_{p+1} = -\frac{f''(\eta_p)}{2f'(\zeta)}\,(\xi - x_p)(\xi - x_{p-1})\,, \tag{3.3.25}$$

for $p = 1, 2, \ldots$, where η_p and ζ_p are between the largest and smallest values of x_{p-1}, x_p, and ξ. If x_{p-1} and x_p are sufficiently close to ξ, if $f'(\xi) \neq 0$, and if $f''(x)$ is bounded in the neighborhood N, then we may assume that

$$\left|\frac{f''(\eta_p)}{2f'(\zeta_p)}\right| \leq M < \infty\,. \tag{3.3.26}$$

Hence, by setting

$$M|\xi - x_P| = \varepsilon_p \tag{3.3.27}$$

we obtain the inequalities

$$\varepsilon_{p+1} \leq \varepsilon_p \varepsilon_{p-1}\ ,\ p = 1, 2, \ldots\,. \tag{3.3.28}$$

Defining $\delta = \max\{\varepsilon_0, \varepsilon_1\}$, ? yields the following inequalities

$$\begin{aligned} \varepsilon_0 &\leq \delta \\ \varepsilon_1 &\leq \delta \\ \varepsilon_2 &\leq \delta^2 \\ \varepsilon_3 &\leq \delta^3 \\ \varepsilon_4 &\leq \delta^5 \\ &\vdots \\ \varepsilon_p &\leq \delta^{m_p} \end{aligned}$$

where

$$\begin{aligned} m_0 &= 1 \\ m_1 &= 1 \\ m_{p+1} &= m_p + m_{p-1} \ , \ p = 1, 2, \dots \ . \end{aligned}$$

The sequence $\{m_p\}_{p=0}^{\infty}$ is called the Fibonacci sequence. It may be shown (see ?) that

$$m_p = \frac{1}{\sqrt{5}} \left(\alpha^{p+1} - \beta^{p+1}\right) \tag{3.3.29}$$

where

$$\alpha = \frac{1+\sqrt{5}}{2} \ , \ \beta = \frac{1-\sqrt{5}}{2} \, . \tag{3.3.30}$$

Thus for large p,

$$m_p \approx \frac{1}{\sqrt{5}} \alpha^{p+1} \approx .44721(1.61803)^{p+1} \tag{3.3.31}$$

and therefore, roughly,

$$|x_{p+1} - \xi| \approx C|x_p - \xi|^{(1+\sqrt{5})/2} \ . \tag{3.3.32}$$

The order of the secant method is thus $q = (1+\sqrt{5})/2 \approx 1.61803$. The number $(1+\sqrt{5})/2$ is called the *golden number*.

The definition of the order q of convergence of a method applies to a system of equations is the same as for scalar equations, except that ? is now replaced by

$$\limsup_{p\to\infty} \frac{||\vec{x}_p - \vec{\xi}||^{q'}}{||\vec{x}_{p+1} - \vec{\xi}||} = \infty \quad \text{for all } q' < q \, . \tag{3.3.33}$$

That is, the order is q if (3.3.33) is satisfied for all $q' < q$.

3.4 Roots of Polynomials

In applications the problem of obtaining a solution for a nonlinear equation often takes the polynomial form

$$p(z) \equiv a_m z^m + a_{m-1} z^{m-1} + \cdots + a_0 = 0 \tag{3.4.1}$$

For reasons associated with a particular application it may be known that the roots of this polynomial are real. For instance this occurs in the case of Gaussian quadrature. At other times the number of real roots is not known a priori. Aside from this question, some applications require knowledge of all the roots while others are only concerned with the root that is largest or smallest in magnitude.

Example 3.4.1 Suppose that an electrical network has a response Y that satisfies the integral equation

$$Y(t) = \int_0^t K(t-u)Y(u)\,du + G(t) \tag{3.4.2}$$

where K and G are given polynomials. Taking the Laplace transform of each side leads to an expression for $y(s)$, the Laplace transform of Y, that is the ratio of two polynomials. Once the roots of the denominator, including multiplicities, are known we can invert $y(s)$ thereby deriving an explicit expression for Y in terms of t.

In this section we consider the approximate solution of (3.4.1) for the variety circumstances mentioned above. The method of bisections usually suffices to quickly produce a rough approximation to a real root. If more accuracy for a real root is required then the secant method is an alternative. Newton's method is appropriate for accurate estimates of complex roots.

3.4.1 Newton's Method for Roots of Polynomials

A positive feature of Newton's method is its easy adaptation to finding both real and complex roots of polynomials. If $p(z)$ is given in the form (3.4.1) we can evaluate $p(z)$ and $q(z) = p'(z)$ simultaneously be means of the following algorithm.

Algorithm 3.4.2

$$\begin{array}{ll} q \leftarrow 0 & \\ p \leftarrow a_m & \\ q \leftarrow zq + p & \\ & n = 1, 2, \ldots, m \\ p \leftarrow zp + a_{m-n} & \end{array}$$

Consider the sequence

$$z_{n+1} = z_n - \frac{p(z_n)}{q(z_n)}\ ,\ n = 0, 1, 2, \ldots \tag{3.4.3}$$

generated by Newton's method. If the coefficients of $p(z)$ are real and the initial guess z_0 is real then necessarily all successive iterates are real. Hence under this scenario it is not possible to find a complex root ξ of $p(z)$. However, for z_0 complex and sufficiently close to ξ the method does indeed converge to ξ.

Example 3.4.3 ?Frank has an example here of finding a real root of a polynomial. Doesn't seem to make sense given the previous paragraph.

3.4.2 Bernoulli's Method

Bernoulli's method is particularly suited for getting the root of largest modulus as well as the root of smallest modulus for (3.4.1). The general solution of the difference equation ? can be written in the form

$$y_n = \sum_{i=1}^{k} \sum_{j=0}^{\mu_i - 1} c_{i,j} n^j z_i^n \tag{3.4.4}$$

where the z_i are the k distinct roots of $p(z)$ with multiplicity μ_i. Hence if $|z_1| > |z_i|$ for $i = 2, 3, \ldots, k$, and if at least one of the $c_{1,j}$ is non-zero, then

$$\frac{y_{n+1}}{y_n} = \frac{\sum_{i=0}^{\mu_1 - 1} c_{i,j}(n+1)^j z_1^{n+1}[1 + o(1)]}{\sum_{j=0}^{\mu_1 - 1} c_{i,j} n^j z_1^n [1 + o(1)]} \tag{3.4.5}$$

approaches z_1 as $n \to \infty$. Thus in this case the difference equation ? is a convenient tool for computing the root of largest modulus.

Since (3.4.4) holds for negative as well as positive integers n, it follows that if $0 < |z_k| < |z_i|$ for $i = k-1, k-2, \ldots, 1$ and if at least one of the $c_{k,j}$ in (3.4.4) is non-zero then by proceeding as in (3.4.5)

$$\frac{y_{n+1}}{y_n} \to \frac{1}{z_k} \text{ as } n \to -\infty\,. \tag{3.4.6}$$

Now suppose that the coefficients of $p(z)$ are real and that the two roots of largest modulus of $p(z)$ are complex conjugate roots. If real values are taken for y_n in (3.4.1) and if $z_1 = \rho e^{i\theta}$, $z_2 = \rho e^{-i\theta}$, then in the case when the roots z_1 and z_2 are simple, $c_1 = ae^{i\alpha}$,$c_2 = ae^{-i\alpha}$, $a > 0$ (3.4.4) yields

$$y_n = 2a\rho^n[\cos(n\theta + \phi) + o(1)] \tag{3.4.7}$$

as $n \to \infty$. Hence for large n,

$$[y_{n+1} - 2\rho\cos\theta y_n + \rho^2 y_{n-1}][1 + o(1)] = 0 \tag{3.4.8}$$

$$[y_{n+2} - 2\rho\cos\theta y_{n+1} + \rho^2 y_n][1 + o(1)] = 0\,. \tag{3.4.9}$$

Solving the pair of equations (3.4.8) and (3.4.9) for ρ and $\cos\theta$ yields

$$\frac{y_n^2 - y_{n+1}y_{n-1}}{y_{n-1}^2 - y_n y_{n-2}} \to \rho^2 \tag{3.4.10}$$

$$\frac{y_{n+1}y_{n-2} - y_n y_{n-1}}{y_{n-2}^2 - y_n y_{n-2}} \to -2\rho\cos\theta \tag{3.4.11}$$

as $n \to \infty$.

Example 3.4.4 Suppose that we want to find the largest root of the cubic polynomial $p(x) = x^3 - 3x^2 + x - 10$ by Bernoulli's method. Then the difference equation we use is

$$y_{n+1} = 3y_n - y_{n-1} + 10y_{n-2}\ ,\ n = 2, 3, \ldots, \tag{3.4.12}$$

where we take $y_0 = y_1 = y_2 = 1$. Table 3.4.1 includes selected values of y_n as well as the ratios y_{n+1}/y_n. Apparently the convergence is very slow. The exact answer, to nine decimals, is 3.522 167 383. Fewer iterates whould have sufficed to yield the same accuracy if we had used Atkin's δ^2 process.

Example 3.4.5 If we apply the Bernoulli method to solve the cubic $p(x) = x^3 - x^2 + 2x + 6 = 0$ then the difference equation becomes

$$y_{n+1} = y_n - 2Y_{n-1} - 6y_{n-2},\ ,\ n = 2, 3, \ldots\ . \tag{3.4.13}$$

Taking $y_0 = y_1 = 0$ and $y_2 = 1$ we calculate the entries in Table 3.4.2 below. In this case $y_{n=1}/y_n$ does not approach a limit, and it is necessary to use (3.4.10) and (3.4.11). The exact answer, to 4 decimals is $\rho = 2.1924$, $\theta = 1.0324$. Notice the very slow convergence.

Table 3.4.1: Bernoulli's method for solving $p(x) = x^3 - 3x^2 + x - 10$.

n	y_n	y_{n+1}/y_n
0	1	1
1	1	1
2	1	12
3	12	3.75
4	45	2.955 555 555
5	133	3.563 909 774
10	74 977	3.526 961 602
15	40 662 788	3.521 973 678
20	22 040 879 700	3.522 171 187
25	11 194 756 442 000	3.522 167 371
26	42 081 321 560 000	

3.4.3 Deflation

If ξ is a root of the polynomial $p(z)$ in (3.4.1) then there is a polynomial $q(z)$ of degree $m-1$ such that

$$p(z) = (z - \xi)q(z)\,. \tag{3.4.14}$$

Setting

$$q(z) = b_{m-1}z^{m-1} + b_{m-2}z^{m-2} + \cdots + b_0 \tag{3.4.15}$$

in (3.4.14) and using (3.4.1), we get

$$\sum_{k=0}^{m} a_k z^k = \sum_{k=0}^{m} (b_k - \xi b_{k-1}) z^k \tag{3.4.16}$$

where $b_m = b_{-1} = 0$. Equating coefficients of equal powers of z on each side yields the recurence relation

$$b_m = 0 \tag{3.4.17}$$

$$b_k = a_{k+1} + \xi b_{k=1}\ ,\ k = m-1, m-2, \ldots, 0$$

or, if we compute the coefficients in reverse order assuming that $\xi \neq 0$,

$$b_{-1} = 0 \tag{3.4.18}$$

Table 3.4.2: Bernoulli's method for solving $p(x) = x^3 - x^2 + 2x + 6 = 0$.

n	y_n	ρ_n^2	$-2\rho_n \cos\theta$	ρ_n	θ
0	0				
1	0				
2	1				
3	1				
4	-1				
5	-9	6.8	-2.2	2.6077	1.1353
10	-101	4.82	-2.30	2.1954	1.0484
15	17 747	4.8947	-2.2463	2.1919	1.0328
16	-7005	4.8114	-2.2500	2.1935	1.0323
17	-101 149	4.8124	-2.2475	2.1937	1.0330
18	-193 621				

$$b_k = \frac{b_{k-1} - a_k}{\xi}, \; k = 0, 1, \ldots, m-1 .$$

Example 3.4.6 Let $p(z)$ be defined as in Example 3.4.1. Take $\xi = z_4$ in Table 3.4.1 and find $q(z)$ be means of (3.4.19). The values of a_k and the compute b_k are given in Table 3.4.3. Thus $q(x) \approx x^2 - 2.248298x + 4.806545$. Notice that since $b_{m-1} = a_m$ the computed b_{m-1} in (3.4.19) provides a check on the accuracy of the previously computed b_k.

Table 3.4.3: The coefficients of $q(x)$ in Example 3.6.5??.

k	a_k	b_k
0	6	4.806 545 277
1	2	-2.248 297 824
2	-1	1.000 000 002
3	1	

Let us next consider the problem of finding all the zeros of the polynomial (3.4.1). In the textbook [12] J. H. Wilkinson has studied the effect of errors in the coefficients of $q(z)$ on the roots of $q(z)$. A small error in a computed root of $p(z)$ may cause a large change in some of the roots of $q(z)$. For example, the polynomial

$$p(z) \equiv (z-1)(z-2)(z-3)\cdots(z-19)(z-20) \tag{3.4.19}$$

obviously has the roots $1, 2, 3, \cdots, 20$. The coefficient of z^{19} in $p(z)$ is -190. If we leave all coefficients of z^k fixed, except for that of z^{19} from which we subtract $2^{-23} \approx 1.19 \times 10^{-7}$, then the exact roots of the resulting polynomial to nine decimals are shown in Table 3.4.4.

Table 3.4.4: Accurate zeros of $(z-1)(z-2)\cdots(z-20) - z^{-23}z^{19}$.

1.00000 0000		6.00000 6944
2.00000 0000		6.99969 7234
3.00000 0000		8.00726 7603
4.00000 0000		8.91725 0249
4.99999 9928		20.84690 8101
10.09526 6145	$\pm$	0.64350 0904i
11.79363 3881	$\pm$	1.65232 9728i
13.99235 8137	$\pm$	2.51883 0070i
16.73073 7466	$\pm$	2.81262 4894i
19.50243 9400	$\pm$	1.94033 0347i

J. H. Wilkinson has shown moreover in [12] that if the relations (3.4.18) are used to repeatedly deflate $p(z)$ then for best numerical accuracy one should compute the zeros in ascending order of magnitude. Similarly if one computes the zero of $p(z)$ that has largest magnitude, then for greatest accuracy in the resulting polynomial $q(z)$ one should use (3.4.19) rather than (3.4.18).

Other direct methods of getting all the roots of $p(z)$ are described in the problems that follow.

It can be shown (see Problem **??**) that if $p(z)$ has roots $z_1, z_2, \ldots, z_m$ and if z_k is a root of p with multiplicity μ, then the polynomial

$$p(x, \varepsilon) \equiv p(z) + \varepsilon r(z) \tag{3.4.20}$$

where

$$r(z) \equiv r_m z^m + r_{m-1} z^{m-1} + \cdots + r_0 \tag{3.4.21}$$

has roots $z_k(\varepsilon)$ where

$$z_k(\varepsilon) - z_k - \left[-\frac{\mu! r(z_k)}{p^{\mu}(z_k)}\right]^{1/\mu} \varepsilon^{1/\mu} = O(\varepsilon^{2/\mu}) \qquad (3.4.22)$$

as $\varepsilon \to 0$. This relation show that a small change in the coefficients of $p(z)$ may cause a relatively large change in the zeros of $p(z)$. For example, if $\mu = 3$ and $\varepsilon = 10^{-9}$ then $varepsilon^{1/\mu} = 10^{-3} = 10^{6} \times 10^{-9}$, i. e. ; the change in z_k may be 10^6 times as large as the change in a coefficient.

Chapter 4

Polynomial Approximations

Introduction

Polynomial computation is an important phase of numerical work. In calculus, a student learns that a polynomial is the simplest type of mathematical function. Thus it may be a surprise to him that polynomial computation presents difficulties and that a careful error analysis is necessary when the polynomial is of moderate or high degree.

In this chapter we derive the Newton and Lagrange polynomials as well as the respective truncation errors. We then address the topic of best approximation to a continuous function by a polynomial of degree n. The norm we choose for error analysis dominates construction of such a polynomial. Two norm options we explore are the L_∞ or Chebyshev norm and the ℓ_2 norm. With respect to the Chebyshev norm we first consider a special case where the approximation leads to the Chebyshev polynomials. We then explicity construct a polynomial of degree n that best approximates a function at $n+2$ select points. Given sufficient continuity conditions the result extends to the entire interval $[a, b]$. With respect to the ℓ_2 norm we seek the best least square approximation on a set of $m > n$ points. This final topic falls into the realm of optimization.

4.1 Properties of Polynomials

We denote an algebraic polynomial of degree n in x as

$$p_n(x) = a_0 + a_1 x + \cdots + a_n x^n \tag{4.1.1}$$

For the present discussion we assume that the a_i are real numbers. If $a_n \neq 0$ then the fundamental theorem of algebra guarantees the existence of n roots x_i where $i = 1, 2, \ldots, n$ such that

$$p_n(x) = a_n(x - x_1)(x - x_2)\cdots(x - x_n) \,. \tag{4.1.2}$$

We do not assume the roots are real nor do we require unicity since the equation above allows for multiplicities.

A powerful consequence of the fundamental theorem of algebra is that if two polynomials of degree n share the same ordinates at $n + 1$ distinct points then the polynomials are identical. To see this, let $p(x)$ and $q(x)$ be polynomials of degree n such that $p(x_i) = q(x_i)$ at x_i where $i = 0, 1, \ldots, n$. The polynomial

$$r(x) = p(x) - q(x)$$

is clearly of degree n and vanishes at $n+1$ distinct points . Using the roots x_i where $i = 1, \ldots, n$ and the fundamental theorem of algebra

$$r(x) = c(x - x_1)\cdots(x - x_n) \,.$$

But $r(x_0)$ also vanishes, hence, $c \equiv 0$.

Problems

4.1.1 Prove that if the coefficients of a polynomial of degree n are real then any complex roots that occur must do so in conjugate pairs. That is, for α and β real if $z = \alpha + i\beta$ is a root of the polynomial then so is $\alpha - i\beta$.

4.1.2 Show that if the roots of a polynomial of degree n are real and the coefficent of x^n is real then necessarily the coefficients of x^j where $j = n-1, \ldots, 0$ are real.

4.2 Newton–Lagrange Interpolation

In this section we introduce interpolatory polynomials. A polynomial $p(x)$ interpolates a function $f(x)$ at $x_0, x_1, \ldots, x_n$ if $p(x_i) = f(x_i)$ for $i = 0, 1, \ldots, n$. Many recipes exist for building $p(x)$, two we examine in detail and others very briefly. We first look closely at the Lagrange polynomial and then develop the Newton polynomial from the Lagrange. Theoretically these represent the same polynomial but, just as with linear systems, alternative constructions lead to different strenths and weaknesses. The discussion below hopefully implies that the Newton polynomial is more practical for computing whereas the Lagrange polynomial is a powerful algebraic tool.

Lagrange recognized the connection between the polynomial that interpolates a function at $n+1$ given points x_i and the $n+1$ polynomials that vanish on the distinct subsets of n points from $\{x_i : \ i = 0, 1, \ldots, n\}$. For instance, polynomial of degree n that vanishes at $x_1, x_2, \ldots, x_n$ must have the form

$$\ell_0(x) = c(x-x_1)(x-x_2)\cdots(x-x_n)\,.$$

Requiring that $\ell_0(x_0) = 1$ determines c and gives the result

$$\ell_0(x) = \frac{(x-x_1)(x-x_2)\cdots(x-x_n)}{(x_0-x_1)(x_0-x_2)\cdots(x_0-x_n)}\,.$$

The same argument produces the polynomial $\ell_i(x)$ that vanishes at $x = x_0, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ and takes the value 1 at $x = x_i$; i.e.,

$$\ell_i(x) = \frac{(x-x_0)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}\,. \tag{4.2.1}$$

In shorthand notation, $\ell_i(x)$ satisfies

$$\ell_i(x_j) = \delta_{ij} \equiv \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j\,. \end{cases} \tag{4.2.2}$$

where i and j run from 0 to n . We refer to the polynomials $\ell_i(x)$ as Lagrange basis functions. If we now multiply $\ell_i(x)$ by $f(x_i)$ and sum over i, we produce the Lagrange interpolating representation

$$L_n(x) = \sum_{i=0}^{n} f(x_i)\ell_i(x)\,. \tag{4.2.3}$$

It follows immediately from the construction argument that $L_n(x) = f(x)$ for $x = x_0, x_1, \ldots, x_n$. Furthermore, the fundamental theorem of algebra implies that $L_n(x)$ is unique. (Why?)

An alternative notation for Lagrange basis functions is useful when we later derive the error associated with Lagrange interpolation. We remark that this notation makes sense if we interpret the quotient x^2/x as x and say that we can extend the domain of a function to include those points where the Taylor series converges. Let $q_{n+1}(x)$ be

$$q_{n+1}(x) = (x - x_0)(x - x_1)\cdots(x - x_n)\,. \tag{4.2.4}$$

Then applying the product rule for derivatives and evaluating the result at x_i yields the denominator of $\ell_i(x)$

$$q'_{n+1}(x_i) = (x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)\,. \tag{4.2.5}$$

This identity gives a new representation for the Lagrange basis functions, namely,

$$\ell_i(x) = \frac{q_{n+1}(x)}{(x - x_i)q'_{n+1}(x_i)}\,. \tag{4.2.6}$$

The Lagrange polynomial $L_n(x)$ now becomes

$$L_n(x) = \sum_{i=0}^{n} \frac{f(x_i)q_{n+1}(x)}{(x - x_i)q'_{n+1}(x_i)}\,. \tag{4.2.7}$$

Note that the Lagrange basis functions provide a convenient algebraic tool for constructing an interpolatory polynomial but they are impractical for calculation. For instance, evaluating (4.2.7) at a point other than an interpolating point is costly. It is possible to simplify $L_n(x)$ with a computer routine but to justify a particular simplification it must lead to fewer calculations that are also stable. Further, a viable simplification must allow for the inclusion of additional interpolatory points without starting from scratch. This directs us toward the Newton polynomial.

We have chosen to derive the Newton polynomial based on the Lagrange. This is historically unjustified but our method nicely shows the connection between the two representations. Observe that $q_{n+1}(x)/(x - x_i)$ is a polynomial of degree n for which the coefficient of x^n is 1. Thus the separate contributions from the Lagrange basis functions that determine the coefficient of x^n in (4.2.7) are easy to identify. If the symbol $f[x_0, x_1, \ldots, x_n]$ represents the coefficient of x^n in $L_n(x)$ then

$$f[x_0, x_1, \ldots, x_n] \equiv \sum_{i=0}^{n} \frac{f(x_i)}{q'_{n+1}(x_i)}\,. \tag{4.2.8}$$

The notation on the left emphasizes that the coefficient of x^n is a function of $x_0, x_1, \ldots, x_n$ but is not meant to convey any particular ordering of points.

In deriving the Lagrange polynomial we made no assumption concerning the order of the x_i. If we relabel the x_i in any order whatsoever, we derive the same polynomial $L_n(x)$; therefore, the value $f[x_0, x_1, \ldots, x_n]$ is independent of ordering. Our subsequent argument relies on this identity.

Suppose that $L_m(x)$ is the Lagrange polynomial of degree m in x, that interpolates $f(x)$ at the points $x_0, x_1, \ldots, x_m$ and that $m = 0, 1, \ldots, n$. By (4.2.8) the coefficient of x^m appearing in $L_m(x)$ is $f[x_0, x_1, \ldots, x_m]$. Now consider $L_m - L_{m-1}$. We make two observations regarding this difference. The first is that the coefficient of x^m must be the same as that appearing in L_m since $L_{m-1}(x)$ is a polynomial of degree at most $m-1$. The second is that $L_m - L_{m-1}$ vanishes at $x_0, \ldots, x_{m-1}$, hence, up to an arbitrary constant c

$$L_m(x) - L_{m-1}(x) = c(x - x_0)(x - x_1)\cdots(x - x_{m-1})\,. \qquad (4.2.9)$$

Combined with our first remark we have

$$c = f[x_0, x_1, \ldots, x_m]\,.$$

If we systematically replace the differences in the identity

$$L_n(x) = L_0(x) + [L_1(x) - L_0(x)] + \cdots + [L_n(x) - L_{n-1}(x)]\,.$$

using (4.2.9) then we have the Newton polynomial

$$\begin{aligned} p_n(x) &= f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &\qquad\qquad (4.2.10) \\ &\quad + \cdots + f[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1)\cdots(x - x_{n-1})\,. \end{aligned}$$

We denote the Newton polynomial by p_n to distinguish the represenation from L_n.

The coefficients $f[\cdots]$ admit convenient calculation. To see why, we first evaluate (4.2.10) at $x = x_1$. This simplifies to

$$p_n(x_1) = f(x_1) = f(x_0) + f[x_0, x_1](x_1 - x_0)$$

so that

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}\,. \qquad (4.2.11)$$

The expression in (4.2.11) is usually called the first divided difference of $f(x)$ at x_0 and x_1. Continuing, if $x = x_2$ in (4.2.10) then

$$\begin{aligned} p_n(x_2) = f(x_2) &= f(x_0) + f[x_0, x_1](x_2 - x_0) \\ &\quad + f[x_0, x_1, x_2](x_2 - x_0)(x_2 - x_1) \end{aligned}$$

or

$$f[x_0, x_1, x_2](x_2 - x_0)(x_2 - x_1) = f(x_2) - f(x_0) - f[x_0, x_1](x_2 - x_0)\,.$$

This gives the second divided difference

$$f[x_0, x_1, x_2] = \frac{f[x_2, x_0] - f[x_0, x_1]}{x_2 - x_1}$$

but the divided difference is symmetric is x_0 and x_1 hence

$$f[x_0, x_1, x_2] = f[x_1, x_0, x_2] = \frac{f[x_2, x_1] - f[x_1, x_0]}{x_2 - x_0}\,.$$

By induction, we deduce the general relation for the m-th divided difference

$$f[x_0, x_1, \ldots, x_m] = \frac{f[x_1, x_2, \ldots, x_m] - f[x_0, x_1, \ldots, x_{m-1}]}{x_m - x_0} \tag{4.2.12}$$

where $f[x_i] \equiv f(x_i)$. In general the first divided difference for any two points $(x_i, f(x_i))$ and $(x_j, f(x_j))$ is given by

$$f[x_i, x_j] \equiv \frac{f(x_i) - f(x_j)}{x_i - x_j} \tag{4.2.13}$$

with higher order divided differences following as above. We can think of $f[x_0, x_1, \ldots, x_m]$ as a linear combination of $f(x_0), f(x_1), \ldots, f(x_m)$ where $x_0, x_1, \ldots, x_m$ determine the exact form of the combination.

Example 4.2.1 Let us find the Lagrange polynomial that passes through the points (0,0), (1,1), and (2,2). Here $q_3(x) = x(x-1)(x-2)$, and so

$$q_3'(0) = (-1)(-2) = 2; \quad q_3'(1) = 1(-1) = -1; \quad q_3'(2) = 2(1) = 2\,.$$

Hence

$$L_2(x) = \frac{0 \cdot q_3(x)}{x(2)} + \frac{1 \cdot q_3(x)}{(x-1)(-1)} + \frac{2 \cdot q_3(x)}{(x-2)(2)} = x\,.$$

Note that, while each polynomial $\ell_i(x)$ of the Lagrange polynomial $L_n(x)$ is of degree $n = 2$ the polynomial $L_n(x)$ may be of lower degree.

Example 4.2.2 The computation of the Newton polynomial is usually carried out according to the following array:

$$\begin{array}{llll}
x_i & y_i & & \\
0 & 1 & & \\
 & & f[x_0,x_1] = \frac{3-1}{1-0} = 2 & \\
1 & 3 & & f[x_0,x_1,x_2] = \frac{7-2}{2-0} = \frac{5}{2} \\
 & & f[x_1,x_2] = \frac{10-3}{2-1} = 7 & \\
2 & 10 & &
\end{array}$$

From this array and (4.2.10) we have $p_2(x) = 1 + 2x + \frac{5}{2}\, x(x-1)$.

The ordering of the points in the array is immaterial. For example,

$$\begin{array}{llll}
x_i & y_i & & \\
2 & 10 & & \\
 & & \frac{1-10}{0-2} = \frac{9}{2} & \\
0 & 1 & & \frac{2-9/2}{1-2} = \frac{5}{2} \\
 & & \frac{3-1}{1-0} = 2 & \\
1 & 3 & &
\end{array}$$

Here we have $p_2(x) = 10 + \frac{9}{2}\,(x-2) + \frac{5}{2}\,(x-2)x$ but this is identical to the first expression.

Note that if the assumption $x_i \neq x_j$ whenever $i \neq j$, is not satisfied, then both the Lagrange method and the Newton method of obtaining an interpolation polynomial fail. We now turn our attention to a selection of topics that arise from general interpolation.

4.2.1 Divided Difference Tables

In the past, the practical use of interpolation polynomials was mainly confined to table–making , an activity of diminishing importance. For convenience tables are usually given at equispaced arguments of the interpolated function.

We begin with Newton polynomials where we assume equispaced points x_i. If h is the tabular interval then the grid points x_i expressed in terms of x_0 and h are

$$x_m = x_0 + mh\,.$$

We reformulate the Newton divided differences by introducing the forward differencing operator Δ. If $f_m \equiv f(x_m)$ then

$$\begin{aligned}
\Delta f_m &\equiv f_{m+1} - f_m \\
\Delta^2 f_m &\equiv \Delta\left(f_{m+1} - f_m\right) = f_{m+2} - 2f_{m+1} + f_m
\end{aligned}$$

and in general

$$\Delta^n f_m = \Delta(\Delta^{n-1} f_m \, .$$

With the definition above, the first and second Newton divided differences are

$$\begin{aligned} f[x_0, x_1] &= \frac{f(x_0) - f(x_1)}{x_0 - x_1} = \frac{f_1 - f_0}{h} = \frac{\Delta f_0}{h} \\ f[x_0, x_1, x_2] &= \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{f_2 - 2f_1 + f_0}{2h^2} = \frac{\Delta^2 f_0}{2!\, h^2} \end{aligned}$$

while induction implies

$$f[x_0, x_1, \ldots, x_m] = \frac{\Delta^m f_0}{m!\, h^m} \, .$$

When these results are substituted into (4.2.10) the Newton polynomial becomes

$$\begin{aligned} p_n(x) &= f_0 + \frac{(x - x_0)}{h} \Delta f_0 + \frac{(x - x_0)(x - x_1)}{2!\, h^2} \Delta^2 f_0 + \cdots \\ &\quad + \frac{(x - x_0)(x - x_1) \cdots (x - x_{n-1})}{n!\, h^n} \Delta^n f_0 \end{aligned}$$

or, in terms of the variable ξ where

$$x = x_0 + \xi h \quad x - x_m = (\xi - m)h$$

$$p_n(x) = \sum_{k=0}^{n} \binom{\xi}{k} \Delta^k f_0 \, . \tag{4.2.14}$$

This last identity points to a disadvantage of equispaced arguments, that is, some coefficients of the resulting polynomial are very large when n is large. For instance, the coefficient of $\Delta^m f_0$ is

$$\binom{\xi}{m} \equiv \frac{\xi(\xi - 1) \cdots (\xi - m + 1)}{m!} \, .$$

The greatest absolute value of this coefficient for ξ in the range $0 \leq \xi \leq n$ is not easy to obtain. For the case $m = 2n + 1$ we find a rough estimate on the magnitude of this coefficient by taking $\xi = n - \frac{1}{2} = 2m + \frac{1}{2}$ and then using Stirling's formula $\Gamma(n+a) \sim (2\pi)^{1/2} n^{n+a-1/2} e^{-n}$ as $n \to \infty$ (see e.g. Chapter 6) to derive

$$\begin{aligned}
\frac{\xi(\xi-1)\cdots(\xi-m+1)}{m!} &= \frac{\left(2m+\frac{1}{2}\right)\left(2m-\frac{1}{2}\right)\cdots\left(m+\frac{1}{2}\right)}{m!} \\
&= \frac{\Gamma\left(2m+\frac{3}{2}\right)}{\Gamma\left(m+\frac{3}{2}\right)\Gamma(m+1)} \\
&= \frac{2^{2m+1}}{\sqrt{\pi}} \frac{\Gamma\left(2m+\frac{3}{2}\right)}{\Gamma(2m+2)} \sim \frac{2^{2m+1}}{(2\pi m)^{1/2}}
\end{aligned}$$

as $m \to \infty$.

The Lagrange polynomial for equispaced abscissae is sometimes useful for mathematical tables. As with the Newton polynomials we expect some simplification in the representation of the Lagrange polynomial. If the interpolating points are

$$x_m = \pm mh \ , \quad m = 0, 1, \ldots, n$$

then the Lagrange basis functions are of degree $2n+1$ with the form

$$\ell_m(x) = \frac{c_m x \prod_{r=1}^{n} (x^2 - r^2 h^2)}{x - mh} \ ,$$

where $c_m = 1/q_{2n+1}(x_m)$, or setting $x = \xi h$

$$\ell_m(x) = c_m h^n \, \frac{\xi \prod_{r=1}^{n} (\xi^2 - r^2)}{\xi - m} \ .$$

Representing $L_{2n+1}(x)$ consists of minor details.

Considering the close link between Lagrange and Newton polynomials we expect large terms in the Lagrange expansion for certain values of $x = \xi h$. To exhibit this we develop a relation between gamma functions and the Langrange basis functions. From Appendix 3, we see that the polynomial

$$\Gamma(\xi + n + 1)/\Gamma(\xi - n)$$

vanishes at the $2n+1$ points $\xi = \pm m$, $m = 0, 1, 2, \ldots, n$; therefore , $(\xi - m)\ell_m(x)$ differs from $\Gamma(\xi + n + 1)/\Gamma(\xi - n)$ by at most a multiplicative constant. If we denote this constant by α then we may write

$$\ell_m(x) = \alpha \, \frac{\Gamma(\xi + n + 1)}{(\xi - m)\Gamma(\xi - n)} \ .$$

Two further identities that we need are

$$\Gamma(x)\Gamma(1-x) = \pi \sin(\pi x)$$

and

$$\sin \pi(\xi - n) = (-1)^{n-m} \sin \pi(\xi - m)\,.$$

Evaluating the first at $x = \xi - n$ and then using the second we now have

$$\ell_m(x) = \alpha(-1)^{n-m}\Gamma(\xi + n + 1)\Gamma(n + 1 - \xi)\,\frac{\sin \pi(\xi - m)}{\pi(\xi - m)}\,.$$

Finally we determine α by requiring that $\ell_m(mh) = 1$ so that

$$\ell_m(x) = \frac{\Gamma(n+1+\xi)\Gamma(n+1-\xi)}{\Gamma(n+1+m)\Gamma(n+1-m)}\,\frac{\sin \pi(\xi - m)}{\pi(\xi - m)}\,.$$

We examine the magnitude of $\ell_m(x)$ for the particular value $\xi = (2n - 1)/2$. In this case

$$|\ell_m((2n-1)h/2)| = \frac{\Gamma\left(2n + \frac{1}{2}\right)\Gamma\left(\frac{3}{2}\right)}{\Gamma(n+1+m)\Gamma(n+1-m)}\,\frac{1}{\pi\left(n - m - \frac{1}{2}\right)}\,.$$

Thus for example,

$$|\ell_0((2n-1)h/2)| = \frac{\Gamma\left(2n + \frac{1}{2}\right)\Gamma\left(\frac{3}{2}\right)}{\Gamma^2(n+1)} \cdot \frac{1}{\pi\left(n - \frac{1}{2}\right)} \sim \frac{2^{2n}}{\pi n}$$

as $n \to \infty$. Other convenient approximation methods exist that do not use equispaced arguments and are free from the disadvantage of large coefficients. Methods based on Chebyshev polynomials are one such instance.

4.2.2 Van der Monde Determinants

The Vandermonde determinant of the form

$$r_2(x) = \begin{vmatrix} 1 & x & x^2 \\ 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \end{vmatrix}$$

arises frequently in polynomial formulations. Clearly, $r_2(x)$ is a polynomial of degree 2 in x that vanishes at $x = x_0$ and at $x = x_1$ since in either of these cases two rows of the determinant are identical. Hence

$$r_2(x) = c_2(x - x_0)(x - x_1)$$

To evaluate c_2 note that the coefficient of x^2 is

$$\begin{vmatrix} 1 & x_0 \\ 1 & x_1 \end{vmatrix}$$

(expansion in minors about row 1). So by uniqueness

$$r_2(x) = (x_1 - x_0)(x - x_0)(x - x_1)\,.$$

Similarly, it may be shown that

$$r_n(x) = \begin{vmatrix} 1 & x & x^2 & \ldots & x^n \\ 1 & x_0 & x_0^2 & \ldots & x_0^n \\ \vdots & & & & \\ 1 & x_{n-1} & x_{n-1}^2 & \ldots & x_{n-1}^n \end{vmatrix} = c_n \prod_{i=0}^{n-1}(x - x_i) \tag{4.2.15}$$

with c_n given by

$$c_n = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^{n-1} \\ 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} \\ \vdots & & & & \\ 1 & x_{n-1} & x_{n-1}^2 & \ldots & x_{n-1}^{n-1} \end{vmatrix} = \prod_{i>j}(x_i - x_j)\,. \tag{4.2.16}$$

The latter equality is obtained by the recursive process in Problem 4.2.6.

If $y(x)$ is the polynomial of degree 2 that passes through the points (x_i, y_i), $i = 0, 1, 2$, then $y(x) = y$ expressed as a functions of the determinant is

$$\begin{vmatrix} y & 1 & x & x^2 \\ y_0 & 1 & x_0 & x_0^2 \\ y_1 & 1 & x_1 & x_1^2 \\ y_2 & 1 & x_2 & x_2^2 \end{vmatrix}\,.$$

The Lagrange polynomial $y(x)$ can be recovered from this equation by expanding the determinant along the first column of the matrix. For example, the coefficient of y_0 in the expansion of the determinant, i.e., $\ell_0(x)$ is given by

$$\ell_0(x) = \begin{vmatrix} 1 & x & x^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{vmatrix} \div \begin{vmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{vmatrix} = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}\,.$$

The condition that $y(x)$ be of degree 2 (and no less) requires that the coefficient of x^2 in the determinant be non–zero, i.e.

$$\begin{vmatrix} y_0 & 1 & x_0 \\ y_1 & 1 & x_1 \\ y_2 & 1 & x_2 \end{vmatrix} \neq 0\,.$$

The geometric interpretation of this result is that the triangle formed by the points $(x_0, y_0), (x_1, y_1), (x_2, y_2)$ does not collapse into a straight line. These results and their geometric interpretations extend easily to the case of a polynomial of degree n.

4.2.3 Hermite Interpolation

A certain type of confluence leads to a Lagrange–type polynomial of degree $2n+1$ that interpolates both f and f' at $x = x_0, x_1, \ldots, x_n$. If we set

$$H_{2n+1}(x) = \sum_{i=0}^{n} [\alpha_i + \beta_i(x - x_i)]\ell_i^2(x)$$

where $\ell_i(x)$ is defined as in (4.2.11), and demand that $H_{2n+1}(x_i) = f(x_i)$, $H'_{2n+1}(x_i) = f'(x_i)$ for $i = 0, 1, \ldots, n$, then we find that

$$\begin{aligned} H_{2n+1}(x) &= \sum_{i=0}^{n} \{f(x_i)[1 - 2(x - x_i)\ell_i'(x_i)]\ell_i^2(x) \\ &\quad + f'(x_i)(x - x_i)\ell_i^2(x)\} \end{aligned} \tag{4.2.17}$$

This is called Hermite's interpolation polynomial and should not be confused with the orthogonal Hermite polynomial that is important in the theory of Gaussian quadrature. Clearly, the above process can be extended. For example we can find the constants a_1, b_1, and c_1 such that the polynomial

$$p_{3n+2}(x) = \sum_{i=0}^{n} [a_i + b_i(x - x_i) + c_i(x - x_i)^2]\ell_i^3(x)$$

satisfies

$$p_{3n+2}^{(k)}(x_i) = f^{(k)}(x_i), \; i = 0, 1, \ldots, n; \quad k = 0, 1, 2\,.$$

Polynomials that interpolate function values plus higher derivatives are generically called osculating polynomials.

4.2.4 Aitken's Variant of the Newton Polynomial

Let us suppose that $n+1$ points $(x_i, f(x_i))$ are given, $i = 0, 1, \ldots, n$, and that we want to evaluate the polynomial $p(\xi)$ of degree n in ξ, at $\xi = x$. We examine

$$P_{0,j}(x) \equiv \frac{\begin{vmatrix} f_0 & x_0 - x \\ f_j & x_j - x \end{vmatrix}}{x_j - x_0}\,, \quad j \neq 0 \tag{4.2.18}$$

and observe that $P_{0,j}(x)$ is the linear polynomial that interpolates f_0 and f_j at x_0 and x_j, respectively. Consider now the quadratic polynomials

$$P_{0,1,j}(x) \equiv \frac{\begin{vmatrix} P_{0,1}(x) & x_1 - x \\ P_{0,j}(x) & x_j - x \end{vmatrix}}{x_j - x_1}\,, \quad j \neq 0, 1$$

that have the property that $P_{0,1,j}(x)$ interpolates f_0, f_1 and f_j at x_0, x_1 and x_j. Now assume that $P_{0,1,2,\ldots,k-1,j}(x)$ has been defined (inductively) to be the polynomial of degree k that interpolates the data $\{f_0, f_1, \ldots, f_{k-1}, f_j\}$ at $\{x_0, x_1, \ldots, x_{k-1}, x_j\}$, $j = k, \ldots, n$, and define

$$P_{0,1,2,\ldots,k-1,k,j}(x) = \frac{\begin{vmatrix} P_{0,1,\ldots,k-1,k}(x) & x_k - x \\ P_{0,1,\ldots,k-1,j}(x) & x_j - x \end{vmatrix}}{x_j - x_k}\,, \quad j \geq k+1\,. \tag{4.2.19}$$

This is the polynomial of degree $k+1$ that interpolates $\{f_0, \ldots, f_k, f_j\}$ at $\{x_0, x_1, \ldots, x_k, x_j\}$, $j = k+1, \ldots, n$. This development is conveniently displayed in tabular form as follows $(k+1 = n = 4)$:

If the columns are enumerated as in the table then the column heading is the degree of the interpolant. The index of $P\ldots$ are the indices of the nodes at which $P\ldots$ interpolates the data. For each fixed x note that an entry in the table is a linear interpolant of two entries in the column to the immediate left, e.g. $P_{0,1,4}$ is the linear interpolant of $P_{0,1}$ and $P_{0,4}$ at x_1 and

Table 4.2.1: Aitken's Difference Table

Column		1	2	3	4
x_0	f_0				
x_1	f_1	$P_{0,1}$			
x_2	f_2	$P_{0,2}$	$P_{0,1,2}$		
x_3	f_3	$P_{0,3}$	$P_{0,1,3}$	$P_{0,1,2,3}$	
x_4	f_4	$P_{0,4}$	$P_{0,1,4}$	$P_{0,1,2,4}$	$P_{0,1,2,3,4}$

x_4 and $P_{0,1,2,3}$ is the linear interpolant of $P_{0,1,2}$ and $P_{0,1,3}$ at x_2 and x_3 (see (4.2.19)).

Problems

4.2.1 If $\ell_m(x)$ is a Lagrange basis function where we are interpolating at the points $x_0, x_1, \ldots, x_n$ show that

$$\sum_{m=0}^{n} \ell_m(x) = 1$$

and

$$\sum_{m=0}^{n} (x_m - x)^j \ell_m(x) = 0 \ , \quad j = 1, 2, \ldots, n \, .$$

4.2.2 Show that

$$f(n) = \sum_{m=0}^{n} \binom{n}{m} \Delta^m f(0) \, .$$

4.2.3 Let $f \in C[a, b]$ and let $x_0, \ldots, x_n$ be points in $[a, b]$. If we define L_n as the operator such that $L_n f$ is a polynomial of degree $\leq n$ which agrees with f at $x_0, \ldots, x_n$ then show that L_n is a linear operator, i.e., $L_n(\alpha f) = \alpha L_n f$ and $L_n(f + g) = L_n f + L_n g$.

4.2.4 Show that $L_n f = f$ if and only if f is a polynomial of degree $\leq n$.

4.2.5 Show that $||L_n f|| \leq \alpha ||f||$ where $\alpha = \max_x \sum |\ell_i(x)|$ and that no smaller α has this property for all $f \in C[a, b]$.

4.2.6 By induction show that

$$c_n = \begin{vmatrix} 1 & x_0 & x_0^2 & \dots & x_0^{n-1} \\ 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ \vdots & & & & \\ 1 & x_{n-1} & x_{n-1}^2 & \dots & x_{n-1}^{n-1} \end{vmatrix} = \prod_{i>j}(x_i - x_j)\,.$$

Reduce c_n to a determinant of order $n-1$ as follows: For columns $i = 1, 2, \dots, n-1$ multiply column i by x_0 and subtract the result from column $i+1$. Then factor and use the induction hypothesis.

4.2.7 Prove that if the $n+1$ numbers x_i, $i = 0, 1, \dots, n$ are distinct, then the system of equations

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \dots & & & & \\ 1 & x_n & x_n^2 & \dots & x^n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y \end{bmatrix}$$

always has a solution for any given set of numbers $y_0, y_1, \dots, y_n$. How is this solution
related to the polynomial of degree n that interpolates the data
$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$? (We do not recommend this as a practical method of constructing an interpolation polynomial.)

4.2.8 Find a_i, b_i, and c_i such that the polynomial $p_{3n+2}(x)$ in (2.20) satisfies $p_{3n+2}(x_i) = f(x_i)$, $p'_{3n+2}(x_i) = f'(x_i)$, $p''_{3n+2}(x_i) = f''(x_i)$, for $i = 0, 1, \dots, n$.

4.2.9 Let $p(x)$ be a polynomial of degree 4 in x, such that $p(0) = 5$, $p(3.5) = 2$, $p(1.7) = 8$, $p(9.6) = 4$, $p(11.2) = -6$. Use the Aitken algorithm to evaluate $p(2.8)$.

4.2.10 Neville's algorithm for evaluating the polynomial $p(x)$ that interpolates $f(x)$ at $x = x_0, x_2, \dots, x_n$ is described as follows:

$$\begin{aligned} p_k^{(0)}(x) &= f(x_k)\,, \quad k = 0, 1, \dots, n \\ p_k^{(s)}(x) &= \frac{1}{x_k - x_{k-s}} \begin{vmatrix} p_k^{(s-1)}(x) & p_{k-1}^{(s-1)}(x) \\ x_{k-s} - x & x_k - x \end{vmatrix} \end{aligned}$$

$k = s, s+1, \dots, n$; $s = 1, 2, \dots, n$. Show that $p_n^{(n)}(x) = p(x)$.

4.2.11 Thiele's algorithm for constructing a continued fraction representation of a rational function $r(x)$ that interpolates $f(x)$ at $x_0, x_1, \ldots, x_n$ takes the form given in the figure below, where the circled term is dropped for purposes of interpolation. The reciprocal divided differences $p(\ldots)$ are defined as follows:

$$\begin{aligned}
p(x_0x_1) &= \frac{x_0 - x_1}{f_0 - f_1} \\
p(x_0x_1x_2) &= \frac{x_0 - x_2}{p(x_0x_1) - p(x_1x_2)} + f_1 \\
p(x_0x_1x_2x_3) &= \frac{x_0 - x_3}{p(x_0x_1x_2) - p(x_1x_2x_3)} + p(x_1x_2) \\
\cdots \\
p(x_0x_1\ldots x_n) &= \frac{x_0 - x_n}{p(x_0x_1\ldots x_{n-1}) - p(x_1x_2\ldots x_n)} + p(x_1x_2\ldots x_{n-1})
\end{aligned}$$

where $f_i = f(x_i)$.

(a) Show that if the circled term in the figure below is retained, then $r(x_i) = f(x_i)$ where $i = 0, 1, \ldots, n$.

(b) Show that if the circled term in the figure is dropped, then $r(x) = \frac{p(x)}{q(x)}$, where p and q are polynomials. Furthermore, if $x, x_0, \ldots, x_n \in [a, b]$,

$$f(x)-r(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_n)}{(n+1)!\,\{q(x)\}^2}\left(\frac{d}{dy}\right)^{n+1}\left[f(y)\{q(y)\}^2\right]_{y=\xi}$$

for some $\xi \in (a, b)$. Notice that

$$f(y) = q^2(y)\left[f(x) - \frac{p(y)}{q(y)}\right] - q^2(x)\left[f(x) - \frac{p(x)}{q(x)}\right]\frac{q_{n+1}(y)}{q_{n+1}(x)}$$

where

$$q_{n+1}(y) = \prod_{i=0}^{n}(y - x_i)$$

has $(n+2)$ zeros and

$$\frac{d^{n+1}}{dy^{n+1}}(qp) = 0\,.$$

where the last two terms of the continued fraction take the form

4.3 The Truncation Error of Interpolating Polynomials

In keeping with the purpose of this chapter, the previous discussion was mainly algebraic. The discussion and the illustrative examples were limited to an understanding of interpolating polynomials, and some extensions of these. The following discussion of the truncation error of polynomial approximation is similarly limited.

Theorem 4.3.1 (Rolle's Theorem): *If*

(i) $f(x)$ *is continuous in* $[a, b]$,

(ii) $f(a) = f(b)$, *and*

(iii) $f'(x)$ *exists in* (a, b),

then there is a point ζ *in* (a, b) *such that* $f'(\zeta) = 0$.

We omit the proof of Rolle's theorem, since it is included in most calculus texts and instead exhibit applications. Our first is the Mean Value Theorem.

Theorem 4.3.2 (Mean Value Theorem): *If* $f(x)$ *is continuous in* $[a, b]$ *and differentiable in* (a, b) *then there exists a point* $\zeta \in (a, b)$ *such that*

$$f'(\zeta) = \frac{f(b) - f(a)}{b - a}.$$

Proof: The linear polynomial that interpolates f at $x = a$ and $x = b$ is

$$p_1(x) = f(a) + f[a, b](x - a).$$

Let

$$\epsilon_1(x) \equiv f(x) - p_1(x),$$

then $\epsilon_1(x)$ satisfies the hypotheses of Rolle's Theorem; hence, there exists $\zeta \in (a, b)$ such that

$$0 = \epsilon_1'(\zeta) = f'(\zeta) - f[a, b].$$

This immediately gives

$$f'(\zeta) = \frac{f(b) - f(a)}{b - a}.$$

One way to interpret the Mean Value Theorem is that it guarantees the existence of at least one point $\zeta \in (a, b)$ such that the tangent line to $f(\zeta)$ is parallel to the secant line between $(a, f(a))$ and $(b, f(b))$. A second interpretation is that among the points $\zeta \in (a, b)$ that satisfy $f'(\zeta) = f[a, b]$ at least one corresponds to $\max\{|\epsilon_1(x)| : a \leq x \leq b\}$. With respect to interpolation our interest is an error estimate that depends more directly on $x \in [a, b]$ and so gives more information for a particular value of x. For linear interpolation this means determining the exact form of $\epsilon_1(x)$. We know that $f(x)$ and $p_1(x)$ match at $x = a$ and $x = b$. To extend the results of the Mean Value Theorem and at the same time use Rolle's Theorem we need another function that produces a zero at any fixed $x \in (a, b)$. One such function is

$$\begin{aligned} g(t) &= [f(t) - p_1(t)](x-a)(x-b) - [f(x) - p_1(x)](t-a)(t-b) \\ &= [f(t) - p_1(t)](x-a)(x-b) - (t-a)(t-b)\epsilon_1(x)\,. \end{aligned}$$

A simple check verifies that $g(t) = 0$ when $t = a, b, x$. Applying Rolle's Theorem once implies that $g'(t)$ vanishes at least twice in (a, b) and a second application gives the existence of $\zeta = \zeta(x)$ such that

$$0 = g''(\zeta) = [f''(\zeta) - p_1''(\zeta)](x-a)(x-b) - 2\epsilon_1(x)\,.$$

But $p_1(x)$ is linear so $p_1''(\zeta(x)) = 0$ for all x and solving for $\epsilon_1(x)$ we have

$$\epsilon_1(x) = \frac{f''(\zeta(x))}{2!}\,(x-a)(x-b)\,.$$

The more general statement we seek follows.

Theorem 4.3.3 *Suppose $f^{(n)}(x)$ is continuous is $[a, b]$ and $f^{(n+1)}(x)$ exists (a, b). Let $p_n(x)$ be a polynomial of degree n in x such that $p_n(x_i) = f(x_i)$ where $i = 0, 1, \ldots, n$ and we assume the x_i are distinct and located in the interval $[a, b]$. Given these conditions, for all $x \in [a, b]$ there exists $\zeta = \zeta(x) \in (a, b)$ such that*

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\zeta(x))}{(n+1)!}\; q_{n+1}(x) \tag{4.3.1}$$

where, as before,

$$q_{n+1}(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$$

Proof: Let us define a function $g(t)$ by

$$g(t) = [f(t) - p_n(t)]q_{n+1}(x) - [f(x) - p_n(x)]q_{n+1}(t)\,, \tag{4.3.2}$$

where x is fixed in $[a, b]$. By induction we can show that $g^{(n+1)}(t)$ vanishes at least once in (a, b), say at $t = \zeta$. We also know that $p_n^{(n+1)}(t) = 0$ for all t and $q_{n+1}^{(n+1)}(t) = (n+1)!$. Differentiating (4.3.1) we conclude that

$$0 = g^{(n+1)}(\zeta(x)) = f^{(n+1)}(\zeta(x))q_{n+1}(x) - [f(x) - p_n(x)](n+1)!\,,$$

or

$$\epsilon_n(x) \equiv f(x) - p_n(x) = \frac{f^{(n+1)}(\zeta(x))}{(n+1)!}\, q_{n+1}(x)\,.$$

4.3.1 The Equispaced Newton Polynomial

We can use the interpolation error to interpret the coefficients of the Newton polynomial. This is easier to do assuming equispaced grid points that also satisfy $x_0 < x_1 < \cdots < x_n$. We begin by noting that the polynomial $q_{n+1}(x)$ simplifies to

$$\begin{aligned} q_{n+1}(x) &= h^{n+1}\xi(\xi-1)\cdots(\xi-n) \\ &= h^{n+1}(n+1)!\binom{\xi}{n+1} \end{aligned}$$

where $x = x_0 + \xi h$ and $\xi \in [0, n]$. If f is an arbitrary function such that $f^{(n+1)}(x)$ exists for $x \in (x_0, x_n)$ then we may write f in terms of the Newton polynomial and error as

$$f(x_0 + \xi h) = \sum_{k=0}^{n} \binom{\xi}{k} \Delta^k f(x_0) + h^{n+1}\binom{\xi}{n+1} f^{(n+1)}(\zeta(x)) \tag{4.3.3}$$

where ζ lies between x_0 and $x_0 + nh$. Similarly, the represenation for f using the Newton polynomial of degree $n-1$ is

$$f(x_0 + \xi h) = \sum_{k=0}^{n-1} \binom{\xi}{k} \Delta^k f(x_0) + h^n \binom{\xi}{n} f^{(n)}(\nu(x)) \tag{4.3.4}$$

where ν lies between x_0 and x_{n-1}. If we now subtract (4.3.4) from (4.3.3), we have

$$\frac{\Delta^n f(x_0)}{h^n} = f^n(\nu(x)) - h\,\frac{\xi - n}{n+1} f^{(n+1)}(\zeta(x))\,.$$

Finally, taking $\xi = n$ and letting $h \to 0$, we find

$$f^{(n)}(x_0) = \lim_{h\to 0} \frac{\Delta^n f(x_0)}{h^n}\,. \tag{4.3.5}$$

If we drop the assumption of equispaced points then we can still arrive at a result similar to the one above. Again we assume that $x_0 < x_1 < \cdots < x_n$. Consider the two expressions

$$\begin{aligned} f(x) &= f(x_0) + f[x_0,x_1](x-x_0) + \cdots \\ &\quad + f[x_0,x_1,\ldots,x_n](x-x_0)(x-x_1)\cdots(x-x_{n-1}) \\ &\quad + \frac{q_{n+1}}{(n+1)!} f^{(n+1)}(\xi(x)) \\ &= f(x_0) + f[x_0,x_1]q_0(x) + \cdots + \\ &\qquad f[x_0,x_1,\ldots,x_n]q_n(x) + \frac{q_{n+1}(x)}{(n+1)!} f^{(n+1)}(\xi(x)) \end{aligned} \tag{4.3.6}$$

and

$$\begin{aligned} f(x) &= f(x_0) + f[x_0,x_1]q_0(x) + \cdots + \\ &\qquad f[x_0,x_1,\ldots,x_{n-1}]q_{n-1}(x) + \frac{q_n(x)}{n!} f^{(n)}(\nu(x))\,. \end{aligned} \tag{4.3.7}$$

Subtracting the second from the first we have

$$\frac{f^{(n)}(\nu(x))}{n!} = f[x_0,x_1,\ldots,x_n] + \frac{x - x_n}{(n+1)!} f^{(n+1)}(\xi(x))\,.$$

Taking $x = x_n$ and letting $x_n \to x_0$ we find that

$$\frac{f^{(n)}(x_0)}{n!} = \lim_{x_n\to x_0} f[x_0,x_1,\ldots,x_n]\,. \tag{4.3.8}$$

Also, by substituting this result into (4.3.6), it follows that

$$f(x) = \lim_{x_n\to x_0} \{f(x_0) + f[x_0,x_1](x-x_0) + \cdots +$$

$$
\begin{aligned}
& f[x_0, \ldots, x_n](x - x_0) \cdots (x - x_n) + \frac{q_{n+1}(x)}{(n+1)!}\, f^{(n+1)}(\zeta(x))\} \\
&&& (4.3.9) \\
= \; & f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}\,(x - x_0)^n \\
& + \frac{f^{(n+1)}(\zeta(x))}{(n+1)!}\,(x - x_0)^{n+1},
\end{aligned}
$$

where $\zeta(x)$ lies between x_0 and x. We therefore see that the Taylor polynomial can be regarded as a special case of the Newton polynomial.

Example 4.3.4 If in (4.3.3) $f(x) = a^x$ and $x_0 = 0$ by induction we can show that $\Delta^k a^{x_0} = (a^h - 1)^k$. Since $f^{(n+1)}(x) = a^x (\log a)^{n+1}$ we now have

$$
a^{\xi h} = \sum_{k=0}^{n} \binom{\xi}{k} (a^h - 1)^k + (h \log a)^{n+1} \binom{\xi}{n+1} a^\nu
$$

where ν is between the largest and smallest of 0, ξh and nh. But

$$
\begin{aligned}
\binom{\xi}{n+1} &= \frac{\xi(\xi - 1)\cdots(\xi - n)}{(n+1)!} \\
&= \frac{(-1)^{n+1}(n - \xi)(n - \xi - 1)\cdots(1 - \xi)(-\xi)}{(n+1)\Gamma(n+1)} \\
&= \frac{(-1)^{n+1}\Gamma(n+1-\xi)}{\Gamma(-\xi)(n+1)\Gamma(n+1)}.
\end{aligned}
$$

Stirling's formula gives

$$
\frac{\Gamma(n+1-\xi)}{\Gamma(n+1)} = \frac{n^{n+1-\xi-1/2}}{n^{n+1-1/2}} = n^{-\xi}
$$

so finally

$$
\left| \binom{\xi}{n+1} \right| = \frac{1}{(n+1)n^\xi \Gamma(-\xi)} \quad \text{as } n \to \infty\,.
$$

However, this last term decays to zero as $n \to \infty$ if $\xi > -1$. If we also assume that $|a^h h \log a| = |h \log a \exp(h \log a)| < 1$ then

$$
\sum_{k=0}^{n} \binom{\xi}{k} (a^h - 1)^k \to a^{\xi h} \tag{4.3.10}
$$

as $n \to \infty$.

The need for care in the practical use of interpolating polynomials has already been suggested in the previous section. Although electronic computers easily evaluate polynomials of high order, it is nevertheless necessary to examine closely the error. One criterion for estimating the error of interpolation is given by (4.3.2). The error is a product of $f^{(n+1)}(\xi(x))/(n+1)!$ and $q_{n+1}(x)$. In the next section we thoroughly examine the dependence of the error on $q_{n+1}(x)$. For example, if $|f^{(n+1)}(x)| \leq M < \infty$ for all x in $[a, b]$, then $|f^{(n+1)}(\xi(x))| \leq M$ for all x in $[a, b]$, and therefore

$$|f(x) - p_n(x)| \leq \frac{M}{(n+1)!} \, |q_{n+1}(x)| \, . \tag{4.3.11}$$

The maximum value of the bound on the right is minimized if we choose $q_{n+1}(x)$ so that its maximum value on the interval $[a, b]$ is a minimum. Specifying $q_{n+1}(x)$ refers to selecting the points $x_0, x_1, \ldots, x_n$. We discuss the minimizations of the maximum of $|q_{n+1}(x)|$ in the following section.

For now, we comment on the effect of the error term $f^{(n+1)}$. To a great extent this term indicates how closely the polynomial preserves the character of the interpolated function. We see that at times polynomials are inadequate tools of approximation. For instance, suppose that $f(x) = 0$ for $x = x_1$, x_2 and x_3, and that $f(x_0) = y_0$ where $x_0 = 0$ and $x_0 < x_1 < x_2 < x_3$. The value y_0 may or may not be finite. If $f(x)$ is unbounded at $x = x_0$ then y_0 simply distinguishes between $\pm\infty$. The polynomial that interpolates $f(x)$ at these four points is

$$p_3(x) = \frac{-y_0(x - x_2)(x - x_3)(x - x_4)}{x_2 x_3 x_4} \, .$$

This polynomial is shown as the dotted curve in Figure 4.3.1.
The polynomial $p_3(x)$ is a good approximation to $f(x)$ if $f(x)$ admits close approximation by a polynomial throughout the interval $[0, x_4]$. This assumption is false when $f(x)$ is infinite at $x = x_1$. In this case no polynomial provides a satisfactory approximation. The rational function

$$g(x) = \frac{-y_0(x - x_2)(x - x_3)(x - x_4)}{x x_2 x_3 x_4} = \frac{p_3(x)}{x}$$

may now be a better approximation. The function $g(x)$ is represented by the continuous curve in Figure 4.3.1. Of course, if we know the rate at which $f(x)$ blows up at $x = x_0$ we can build this information into the rational approximation. For example, if

$$f_1(x) = \frac{\sin x}{x^2} \quad \text{and} \quad f_2(x) = \frac{\sin x}{x^3}$$

Figure 4.3.1: Example of Interpolation

then the rational interpolant for $f_1(x)$ should behave like $1/x$ near $x = 0$ while the interpolant behavior for $f_2(x)$ should mimic $1/x^2$ near $x = 0$.

This example is not artificial; rather, it emphasizes an essential limitation of polynomial approximation. Several excellent references are given at the end of this chapter. We note that the theory of the Gamma function is the most completely worked out theory of an interpolating function.

Problems

4.3.1 Verify (4.3.10).

4.3.2 *(This problem should be performed with the aid of a computer package such as MATLAB.)* Consider the function $f(x) = 1/(1+25x^2)$ on the interval $[-1, 1]$. Interpolate $f(x)$ at the points $x_i = -1 + ih$ where $h = .2$ and $i = 0, 1, \ldots, 11$. Graph $f(x)$ with a solid line and the interpolant with a dashed line. Comment on the quality of interpolant. Extend the interval to $[-2, 2]$ and perform a similar analysis.

4.3.3 Prove that the polynomial $H_{2n+1}(x)$ in (4.2.17) satisfies

$$f(x) - H_{2n+1}(x) = \frac{f^{(2n+2)}(\xi(x))}{(2n+2)!}\,(x - x_0)^2(x - x_1)^2 \cdots (x - x_n)^2$$

where $\xi(x)$ is between the largest and smallest of $x_0, x_1, \ldots, x_n$ and x.

4.3.4 (a) Show that $f[x,x] = \dfrac{d}{dx}f[x]$. Hint: Let $x_1 = x+\delta$ and consider $\lim_{\delta\to 0} f[x_1, x]$.

(b) Show that $f[x_0, x_1, \ldots, x_m, x, x] = \dfrac{d}{dx}f[x_0, x_1, \ldots, x_m, x]$.

4.3.5 Theorem 4.3.3 gives us the exact from of $\epsilon_1(x)$ where

$$f(x) = f[x_0] + (x - x_0)f[x_0, x_1] + \epsilon_1(x)\,.$$

Using the relation above and the identity

$$f[x_0, x_1, x] = \frac{f[x_0, x] - f[x_0, x_1]}{x - x_1}$$

show that

$$f[x_0, x_1, x] - \frac{f''(\zeta(x))}{2!}\ ,\quad \zeta \in (x_0, x_1)\,.$$

In general show that if $x_0, x_1, \ldots, x_n$ lie in $[a, b]$ then

$$f[x_0, x_1, \ldots, x_n, x] = \frac{f^{(n+1)}(\zeta(x))}{(n+1)!}\ ,\quad \zeta \in (a, b)\,.$$

4.3.6 Show that if $f(x) = \mathrm{erf}x$ where

$$\mathrm{erf}x \equiv \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt$$

then $|\epsilon_1(x)|$ cannot exceed $(x_1 - x_0)^2/(2\sqrt{2\pi e})$ between x_0 and x_1.

4.3.7 Suppose that $f(x) = \int_0^x \sin t^2\,dt$ is tabulated for $0 \le x \le 1$ with uniform spacing h. What spacing should be used to guarantee an accuracy within 0.005?

4.3.8 Let $x_0, x_1, \ldots, x_n$ lie in the interval $[a, b]$. If $f(z)$ is analytic in a simply connected region D fo the complex plane which includes the line segment $[a, b]$ and if C is a simple closed curve in D that surrounds $[a, b]$ show that

$$f[x_0, x_1, \ldots, x_n, x] = \frac{1}{2\pi i}\oint_C \frac{f(z)\,dz}{q_{n+1}(z)(z - x)}$$

where, as usual, $q_{n+1}(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$. Then show that

$$|\epsilon_n(x)| \le \frac{ML}{2\pi d}\max_{z\in C}\left|\frac{q_{n+1}(x)}{q_{n+1}(z)}\right|$$

where $|f(z)| \le M$ on C, L is the length of C and d is the shortest distance from the real point x to a point on C.

4.4 The Polynomial of Least Deviation from Zero in $[-1,1]$

In the previous section we derived the error statement

$$f(x) - p_n(x) = \epsilon_n(x) \equiv \frac{f^{(n+1)}(\zeta(x))q_{n+1}(x)}{(n+1)!} \tag{4.4.1}$$

assuming that $x, x_i \in [a,b]$, and $f^{(n+1)}(x)$ exists for all $x \in (a,b)$. In this case $\zeta(x) \in (a,b)$ if $x \in [a,b]$; therefore we can bound $\epsilon_n(x)$ if $f^{(n+1)}(x)$ is bounded on $[a,b]$ as follows:

$$|\epsilon_n(x)| \le \left(\max_{x\in[a,b]} \left|\frac{f^{(n+1)}(x)}{(n+1)!}\right|\right)\left(\max_{x\in[a,b]} |q_{n+1}(x)|\right) . \tag{4.4.2}$$

We make two comments regarding this approach. The first is that we have implicity chosen to discuss error in terms of the infinity norm. The second is that separating the maximization of $|f^{(n+1)}(x)|$ from that of $|q_{n+1}(x)|$ is not the only way to proceed. The next two sections offer alternatives to these options. The proper choice of norm depends on a given application and more than one choice is often valid. We do not say more in this regard. However, with respect to segregating the minimization of $\max|q_{n+1}(x)|$ we see the advantage of a more universally applicable result since a reasonable small value for this term is satisfactory for the interpolation of many functions.

We begin by translating the abstract problem of minimizing the $\max|q_{n+1}(x)|$ to a concrete problem with a list of criteria. Recall that

$$q_{n+1}(x) = (x - x_0)(x - x_1)\cdots(x - x_n) .$$

One feature this polynomial has independent of x_i is that the coefficient of x^{n+1} is $a_{n+1} = 1$. We now define the minimum deviation problem as finding the polynomial of degree $n+1$ whose leading coefficient is 1, whose zeros $x_0, x_1, \ldots, x_n$ lie in $[a,b]$, and such that its uniform norm on $[a,b]$ is as small as possible. We must select the proper set of zeros to complete the minimax problem. Rather than perform the analysis on $[a,b]$ we choose to work on $[-1,1]$ and then explain how to transfer the result to any finite interval $[a,b]$. Hence we are now assuming $x \in [-1,1]$.

We show that the solution of

$$\min_{x_i\in[-1,1]} \max|q_{n+1}(x)|$$

is satisfied by the choice

$$q_{n+1}(x) = t_{n+1}(x)$$

where

$$t_n(x) = \frac{\cos n\theta}{2^{n-1}} \quad , \quad x = \cos\theta \quad , \quad \theta \in [0, \pi] \tag{4.4.3}$$

is known as the n-th *Chebyshev* polynomial of the second kind. Several comments are in order. First the mapping $x = \cos\theta$ is one–to–one from $[0, \pi]$ onto $[-1, 1]$. Next, $t_n(x)$ is indeed a polynomial. To prove this note that

$$\begin{aligned} t_0(x) &= 2 \\ t_1(x) &= x \,. \end{aligned}$$

Assuming that $t_j(x)$ where $j = 2, \ldots, n$ are polynomials and using the trigonometric identity

$$\cos(a + b) + \cos(a - b) = 2\cos a \cos b$$

gives

$$\cos(n + 1)\theta = 2\cos n\theta \cos\theta - \cos(n - 1)\theta$$

or

$$t_{n+1}(x) = xt_n(x) - \frac{1}{4}\, t_{n-1}(x) \,.$$

Hence, by induction we have shown that $t_n(x)$ is a polynomial of degree n whose leading coefficient is $a_n = 1$. We also have a recursion relation so that, for instance,

$$\begin{aligned} t_2(x) &= xt_1(x) - \tfrac{1}{4}t_0(x) &&= x^2 - \frac{1}{2} \\ t_3(x) &= x\left(x^2 - \tfrac{1}{2}\right) - \tfrac{1}{4}\,x &&= x^3 - \frac{3}{4}\,x \,. \end{aligned}$$

Next, we must examine the zeros and critical points of $t_{n+1}(x)$. The former correspond to

$$\theta_k = \frac{(2k + 1)\pi}{2(n + 1)} \quad , \quad k = 0, 1, \ldots, n \tag{4.4.4}$$

and the latter critical points occur when

$$t'_{n+1}(x) = \frac{n + 1}{2^n(1 - x^2)^{1/2}}\, \sin(n + 1)\theta = 0 \,.$$

To distinguish between the zeros of $t_{n+1}(x)$ and the critical points we denote the critical points as $\hat{\theta}_k$ where, from the equation above,

$$\hat{\theta}_k = \frac{k\pi}{n+1}\,, \quad k = 0, 1, \ldots, n\,. \tag{4.4.5}$$

Since $|\cos(n+1)\hat{\theta}_k| = 1$ for all k we have

$$\max_{-1\le x\le 1} |t_{n+1}(x)| = 2^{-n}\,. \tag{4.4.6}$$

Suppose that we interpolate a function f on the interval $[-1,1]$ at the points $x_i = \cos\left[\frac{(2i+1)\pi}{2n+2}\right]$, $i = 0, 1, \ldots, n$ using the Lagrange representation. In this case

$$\ell_k(x) = \frac{t_{n+1}(x)}{(x-x_k)[t'_{n+1}(x)]\,|_{x=x_k}}\,.$$

Now

$$\frac{d}{dx}\cos(n+1)\theta = \frac{(n+1)\sin(n+1)\theta}{(1-x^2)^{1/2}}$$

and so

$$\frac{d}{dx}\cos(n+1)\theta\,|_{\theta=\theta_k} \;=\; \frac{(n+1)(-1)^k}{(1-x_k^2)^{1/2}}\,.$$

Substituting into the expression for $\ell_k(x)$ gives

$$\ell_k(x) = \frac{(-1)^k(1-x_k^2)^{1/2}2^n t_{n+1}(x)}{(n+1)(x-x_k)}\,. \tag{4.4.7}$$

¿From the results above we now have

$$\left| f(x) - \sum_{k=0}^{n} \frac{f(x_k)(-1)^k(1-x_k^2)^{1/2}2^n t_{n+1}(x)}{(n+1)(x-x_k)} \right| \le 2^{-n} \left| \frac{f^{(n+1)}(\zeta(x))}{(n+1)!} \right|\,. \tag{4.4.8}$$

It may also be shown that (Problem 4.4.3).

$$|\ell_i(x)| \le 2 \;\text{ if }\; |x| \le 1\,. \tag{4.4.9}$$

Let us now state and prove a theorem that characterizes the polynomial $t_{n+1}(x)$.

Theorem 4.4.1 *Let*

$$p_{n+1}(x) = x^{n+1} + a_n x^n + \cdots + a_0 \tag{4.4.10}$$

and let us set

$$G(p_{n+1}) = \max_{-1\le x\le 1} |p_{n+1}(x)|\,. \tag{4.4.11}$$

Then $G(p_{n+1}) \ge 2^{-n}$. *Equality occurs if and only if*

$$p_{n+1}(x) = t_{n+1}(x) = 2^{-n}\cos((n+1)\arccos x)\,.$$

Proof: (by contradiction) Suppose that there is a polynomial $p_{n+1}(x)$, say

$$p_{n+1}(x) = x^{n+1} + b_n x^n + \cdots + b_0$$

such that

$$G(p_{n+1}) < 2^{-n}\,.$$

Then

$$r_n(x) = p_{n+1}(x) - t_{n+1}(x)$$

is a polynomial of degree n since both coefficients of x^{n+1} in $p_{n+1}(x)$ and $t_{n+1}(x)$ are 1. Furthermore $r_n(x)$ is not identically zero. At the points $x_k = \cos(\hat\theta_k)$ where $k = 0, 1, \ldots, n$

$$t_{n+1}(x_k) = (-1)^k 2^{-n}$$

but by hypothesis

$$|p_{n+1}(x_k)| < 2^{-n}\,.$$

Therefore $r_n(x_0) < 0$, $r_n(x_1) > 0, \ldots, r_n(x_k)(-1)^k < 0, \ldots$ or, more simply, $r_n(x)$ changes sign $n+1$ times on the interval $(-1,1)$. Now the polynomial $r_n(x)$ is only of degree n in x, hence $r_n(x) \equiv 0$, a contradiction. This gives $G(p_{n+1}) \ge 2^{-n}$ but equality follows immediately from the construction of $t_{n+1}(x)$.

We close this section by transferring our results to the interval $[a,b]$. The interval $-1 \le \xi \le 1$ maps one–to–one onto the interval $a \le x \le b$ via

$$x = \frac{a+b}{2} + \frac{b-a}{2}\,\xi\,.$$

If $x_i = (a+b)/2 + (b-a)\xi_i/2$ then

$$x - x_i = \frac{b-a}{2}\,(\xi - \xi_i)$$

and

$$q_{n+1}(x) = \left(\frac{b-a}{2}\right)^{n+1} \prod_{i=0}^{n} (\xi - \xi_i) .$$

¿From the analysis above we know that $\max\{|q_{n+1}(x)| : a \le x \le b\}$ is minimized by selecting

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos\theta_k$$

where $\theta_k = \cos((2k+1)\pi/(2n+2))$. With this set of zeros we denote $q_{n+1}(x)$ by $\hat{t}_{n+1}(x)$ and note that

$$\begin{aligned} \max_{a \le x \le b} |\hat{t}_{n+1}(x)| &= \left(\frac{b-a}{2}\right)^{n+1} max_{-1 \le \xi \le 1} |t_{n+1}(x) \\ &= \left(\frac{b-a}{2}\right)^{n+1} 2^{-n} \\ &= \frac{(b-a)^{n+1}}{2^{2n+1}} . \end{aligned}$$

Problems

4.4.1 Show that $t_n(x)$ is even or odd respectively when n is even or odd.

4.4.2 The polynomial $T_n(x) = \cos n\theta$, where $x = \cos\theta$, is called the Chebyshev polynomial of degree n in x. Clearly

$$T_n(x) = 2^{n-1} t_n(x).$$

Show that the Chebyshev polynomials have the following properties.

(a) $(1-x^2)T_n''(x) - xT_n'(x) + n^2 T_n(x) = 0$

(b) $T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x)$ where $n = 2, 3, \ldots$

(c) Orthogonality

$$\int_{-1}^{1} T_n(x) T_m(x) \frac{dx}{(1-x^2)^{1/2}} = \begin{cases} 0 & n \ne m \\ \pi & n = m = 0 \\ \pi/2 & n = m \ne 0 \end{cases}$$

(d) Semi-group property

$$T_m(T_n(x)) = T_{mn}(x) \ ; \quad m, n > 0$$

4.4.3 Show that if

$$x_i = \cos\left[\frac{(2i+i)\pi}{2(n+1)}\right] \quad , \quad i = 0, 1, \ldots, n \, ,$$

and

$$\ell_i(x) = \frac{(-1)^i (1 - x_i^2) \cos[(n+1)\arccos x]}{(n+1)(x - x_i)}$$

then $|\ell_i(x)| \le 2$ for $-1 \le x \le 1$.

4.4.4 Let $p_n(x)$ be the polynomial of degree n in x that interpolates e^x at the zeros of $T_n(x) = \cos(n \arccos x)$. Prove that for $-1 \le x \le 1$,

$$|e^x - p_n(x)| \le \frac{e}{2^n n!} \, .$$

4.5 Best Approximation

The theory and background of best approximation is a deep subject that evolved over many years. We can not duplicate this work in one section but we can and do highlight the major results.

Let f be a given continuous function on the interval $[-1, 1]$. We seek a polynomial $p_{n+1} \in \Pi_{n+1}$ that approximates f well. If by "approximating well" we mean exact interpolation of f at the $(n+2)$ points $\{x_i\}_{i=0}^{n+1}$ where $-1 \le x_0 < x_1 < \cdots < x_{n+1} \le 1$ then a solution of the problem is provided by the Lagrange interpolate L_{n+1} of f at $\{x_i\}$. However, the error between f and L_{n+1} away from the nodes can be huge. If L_{n+1} is used in place of f in a numerical process such as integration then L_{n+1} may not approximate f well. We must always specify the approximation considered as well as the norm measuring error. Throughout this chapter we examine only interpolation via polynomials. In the next section we measure error in terms of the L_2 norm while in this section we turn our attention to the L_∞ norm.

Before stating the precise definition of best approximation in the L_∞ norm we motivate the discussion with the following simple situation. Find $b_1(x) = a_1 x + a_0$ so that for $f \in C[-1, 1]$

$$E_1(f) = \max_{-1 \le x \le 1} |f(x) - b_1(x)| \le \max_{-1 \le x \le 1} |f(x) - p(x)| \tag{4.5.1}$$

for all $p \in \Pi_1$. In minimax language define

$$E_1(f) = \max_{-1 \leq x \leq 1} |f(x) - b_1(x)| = \min_{p \in \Pi_1} \left\{ \max_{-1 \leq x \leq 1} |f(x) - p(x)| \right\}$$

where for now we assume the problem is well defined and $b_1(x)$ exists. The solution for a given f is justifiable geometrically. Referring to Figure 4.5.1 note that the dotted line corresponding to $d_1(x) = \alpha x + \beta$ cannot be the solution since by lowering the line

$$|f(x) - \tilde{d}_1(x)| < |f(x) - d_1(x)| \,.$$

On the other hand, the solid curve $b_1(x)$ looks to be a better approximate

Figure 4.5.1:

to f than either d_1 or $\tilde{d}_1$. We eventually show that $b_1(x)$ is the unique best approximate for the problem defined by (4.5.1). The distinguishing feature $b_1(x)$ exhibits is that the error $E_1(f)$ occurs at three points and with alternate signs. Hopefully, this provides some motivation for the general setting outlined in the following two theorems.

Theorem 4.5.1 *Let $f \in C[a,b]$. Then there exists a unique $b_n \in \Pi_n$ satisfying*

$$E_n(f) \equiv \max_{x \in [a,b]} |f(x) - b_n(x)| \leq \max_{x \in [a,b]} |f(x) - p_n(x)| \qquad (4.5.2)$$

for all $p_n \in \Pi_n$*. Moreover there are at least* $n+2$ *points* $x_0 < x_1 < \cdots < x_{n+1}$ *in* $[a,b]$ *so that*

$$f(x_i) - b_n(x_i) = (-1)^i E_n(f)\,,\; i = 0, \ldots, n+1\,. \tag{4.5.3}$$

The set $\{x_i\}_{i=0}^{n+1}$ in Theorem 4.5.1 is called an alternating set for the function f and the polynomial b_n is called the best approximate to f from Π_n. The "equi–oscillation" of the error in (4.5.3) characterizes best approximates. We emphasize that while b_n is unique this is not necessarily true of the alternate set.

Theorem 4.5.2 *Assume that* $f \in C[a,b]$ *and there is a* $p_n \in \Pi_n$ *satisfying*

$$\max_{a \le x \le b} |f(x) - p_n(x)| = \mu\,.$$

If there is a set of $n+2$ *points* $a \le x_0 < x_1 < \cdots < x_{n+1} \le b$ *so that*

$$f(x_i) - p_n(x_i) = (-1)^i \mu\,, \quad i = 0, 1, \ldots, n+1$$

then $\mu = E_n(f)$ *and* $p_n \equiv b_n$*.*

To gain insight into the previous theorems and what is involved in constructing a best approximate, consider the following example.

Example 4.5.3 The function $f(x) = |x|$ is in $C[-1,1]$. Suppose we seek the best quadratic approximate b_2 to f. Convince yourself that $b_2(x) = \alpha x^2 + \beta$ since f is even; indeed, for the same reason $b_3(x) = b_2(x)$. An alternation set for f contains at least $n+2 = 4$ points but due to symmetry actually contains 5 points. Label these $\{-1, -\xi, 0, \xi, 1\}$ (Problem 4.5.1 asks you to show that 0 and ± 1 belong to the alternation set). Using (4.5.3) we have

$$\begin{aligned} f(-1) - b_2(-1) = 1 - (\alpha + \beta) &\quad = -E_1(f) && \text{(a)} \\ f(-\xi) - b_2(-\xi) = \xi - (\alpha\xi^2 + \beta) &\quad = +E_1(f) && \text{(b)} \\ f(0) - b_2(0) = -\beta &\quad = -E_1(f) && \text{(c)} \end{aligned}$$

Equation (c) gives $\beta = E_1(f)$; then using this result along with equation (a) we have $\alpha = 1$. The alternating set determines the value of β. We know that ξ corresponds to a critical point of (b). Differentiating equation (b) with respect to ξ shows that $\xi = 1/2$. Finally, solving (b) given α and ξ we have $\beta = 1/8$. Hence $b_2(x) = x^2 + 1/8$.

The previous example is typical. That is, the $n+2$ equations in (4.5.3) defining $E_n(f)$ and the $n+1$ coefficients of b_n are nonlinear in the sense that we must determine an alternating set. Theorem 4.5.2 is a basis for approximation schemes that compute $b_n(x)$ but if a high degree approximation is

Figure 4.5.2: Best Approximation to $f(x) = |x|\ , \quad -1 < x < 1\,.$

sought then determining an alternating set valid on the interval $[a, b]$ is often prohibitively expensive. Hence we defer further discussion of this topic to the various approximation theory books of the appendix and instead consider other options. One option is that we split $[a, b]$ into subintervals and calculate low degree approximations on each. This is the basis of splines. The second option is that we seek a best approximation on a set of points in $[a, b]$ rather than on the entire interval. This is the subject of the next two theorems.

Theorem 4.5.2 remains valid if f is defined at $n+2$ points $x_0 < x_1 < \cdots < x_{n+1}$ and the maximum is taken over this finite set. On the finite set we can explicity calculate the best approximation and the error. The disadvantage is that the error may not extend to the entire interval. The theorem that covers this setting is easier to state if we first construct the best approximate on the set of points. This construction combines the ideas of Lagrange interpolation and alternating sets. Note at the conclusion that the algorithm is linear. Let S_{n+2} denote a finite set of $n+2$ points (x_i, y_i),

$i = 0, 1, \ldots, n+1$. We seek the unique polynomial $b_n^*(x) \in \Pi_n$ such that

$$E_n^*(f) = \max_{(x_i,y_i)\in S_{n+2}} |y_i - b_n^*(x_i)| \le \min_{p_n\in\Pi_n} \left\{ \max_{(x_i,y_i)\in S_{n+2}} |y_i - p_n(x_i)| \right\} . \tag{4.5.4}$$

Recall that the Lagrange polynomial that interpolates the function values $y_i = f(x_i)$ for $i = 0, 1, \ldots, n = 1$ is

$$L_{n+1}(x) = \sum_{i=0}^{n+1} \frac{y_i q_{n+2}}{(x - x_i) q'_{n+2}(x_i)}$$

where

$$q_{n+1}(x) = \prod_{i=0}^{n+1} (x - x_i) .$$

Clearly $L_{n+1}(x)$ is not the polynomial we seek since $L_{n+1}(x) \in \Pi_{n+1}$ and we have not yet introduced an alternating set. However, a sketch of $q_{n+2}(x)$ and $q'_{n+2}(x)$ suggests a way to introduce such a set. This polynomial has the properties that $q'_{n+2}(x_i) \neq 0$ and $(-1)^i q'_{n+2}(x_i)$ has the same sign for $i = 0, 1, \ldots, n+1$. The first property gives a nonzero scaling factor while the second guarantees points of alternation so that the polynomial

$$u_{n+1}(x) = \sum_{i=0}^{n+1} \frac{(-1)^i q_{n+2}(x)}{(x - x_i) q'_{n+2}(x_i)}$$

satisfies the identity $u_{n+1}(x_i) = (-1)^i$. Further, the sign property implies that the coefficient of x^{n+1} of $u_{n+1}(x)$ does not vanish. Hence, in general both $L_{n+1}(x)$ and $u_{n+1}(x)$ are polynomials of degree $n+1$. The coefficient of x^{n+1} in $L_{n+1}(x)$ is

$$\sum_{i=0}^{n+1} \frac{y_i}{q'_{n+2}(x_i)}$$

while the corresponding coefficient for $u_{n+1}(x)$ is

$$\sum_{i=0}^{n+1} \frac{(-1)^i}{q'_{n+2}(x_i)} .$$

Given that we know the coefficients of x^{n+1} an explicit method exists for combining $L_{n+1}(x)$ and $u_{n+1}(x)$ so that

$$b_n^*(x) = L_{n+1}(x) - \mu u_{n+1}(x) \tag{4.5.5}$$

where

$$\mu = \sum_{i=0}^{n+1} \frac{y_i}{q_{n+1}(x_i)} \Big/ \sum_{i=0}^{n+1} \frac{(-1)^i}{u'(x_i)} . \tag{4.5.6}$$

We have not shown that this combination actually gives $b_n^*(x)$ but the statement is the next theorem.

Theorem 4.5.4 *The polynomial $b_n^*(x)$ defined in (4.5.5) and (4.5.6) is the unique polynomial that satisfies (4.5.4) with $m = n + 2$.*

In the case of a set S_m of m points (x_i, y_i), where $m \geq n + 2$, and where we assume that the x_i are all distinct. From S_m we can select $\binom{m}{n+2}$ different sets $S_{m,k}$, $k = 1, 2, \ldots, \binom{m}{n+2}$, where

$$S_{m,k} = \left\{ \left(x_0^{(k)}, y_0^{(k)}\right), \left(x_1^{(k)}, y_1^{(k)}\right), \ldots, \left(x_{n+1}^{(k)}, y_{n+1}^{(k)}\right) \right\}$$

and where $x_0^{(k)} < x_1^{(k)} < \cdots < x_{n+1}^{(k)}$. On each of the sets $S_{m,k}$ we approximate the ordinates by a polynomial $p_{n,k}^*(x)$ chosen from Π_n and constructed along the lines following (4.5.4) such that

$$y_j^{(k)} - p_{n,k}^* \left(x_j^{(k)}\right) = (-1)^{j+1} \mu_k \ , \quad j = 0, 1, \ldots, n + 1 \, . \tag{4.5.7}$$

Let ℓ be the integer chosen from $\left\{1, 2, \ldots, \binom{m}{n+2}\right\}$ such that $|\mu_k| \leq |\mu_\ell|$ for $k = 1, 2, \ldots, \binom{m}{n+2}$, and let $p_{n,\ell}^*(x)$ be the corresponding polynomial be the polynomial corresponding to the set $S_{m,\ell}$.

Theorem 4.5.5 *Let $p_{n,\ell}^*$ be the polynomial defined as in the above paragraph. Then $p_{n,\ell}^*(x)$ is the unique polynomial in Π_n that satisfies (4.5.4). Moreover, on the set of point $\{x_i\}_{i=0}^m$*

$$|y_i - p_{n\ell}^*(x_i)| \leq |\mu_\ell| \, . \tag{4.5.8}$$

Proof: We only prove the uniqueness of $p_{n,\ell}^*(x)$; we illustrate the construction of $p_{n,\ell}^*(x)$ via the exchange algorithm, by means of examples.

Let us assume that there are two distinct polynomials, $p_{n,\ell}^*(x)$ and $\hat{p}_n(x)$, both of which are in Π_n, and satisfy

$$|y_i - p_{n,\ell}^*(x_i)| \leq |\mu_\ell| \ , \quad |y_i - \hat{p}_n(x_i)| \leq |\mu_\ell| \ , \quad i = 0, 1, \ldots, m \, . \tag{4.5.9}$$

We set $r_n(x) = \frac{1}{2} \left[p_{n,\ell}^*(x) + \hat{p}_n(x)\right]$. Then there are again $n+2$ points (x_{s_i}, y_{s_i}) in S_m, $x_{s_0} < x_{s_1} < \cdots < x_{s_{n+1}}$ (why?) such that

$$y_{s_i} - r_n(x_i) = (-1)^{i+1} \mu \ , \quad i = 0, 1, \ldots, n + 1 \, ,$$

where $\mu = \pm|\mu_\ell|$ (why?). That is,

$$[y_{s_i} - p^*_{n,\ell}(x_{s_i})] + [y_{s_i} - \hat{p}_n(x_{s_i})] = 2(-1)^{i+1}\mu\,.$$

However, because of (4.5.9), it follows that

$$y_{s_i} - p^*_{n,\ell}(x_{s_i}) = y_{s_i} - \hat{p}_n(x_{s_i}) = (-1)^{i+1}\mu\,,$$

that is, $p^*_{n,\ell}(x_{s_i}) = \hat{p}_n(x_{s_i})$. Since there are $n+2$ such points, $p^*_{n,\ell}(x) = \hat{p}_n(x)$ for all x.

Example 4.5.6 For given $S_3 = \{(x_i, y_i) : i = 0, 1, 2\}$ let us construct the straight line, $b_1(x) = ax + b$ that satisfies $y_i - p_i(x_i) = (-1)^{i+1}\mu$, $i = 0, 1, 2$. However, instead of using the formulas (4.5.5) to (4.5.6), we carry out the computation directly by solving the linear system

$$\begin{aligned} ax_0 + b &= y_0 - \mu \\ ax_1 + b &= y_1 + \mu \\ ax_2 + b &= y_2 - \mu \end{aligned}$$

which yields

$$\begin{aligned} a &= \frac{y_2 - y_0}{x_2 - x_0} \\ b &= \frac{1}{2}\,\frac{y_0(x_1 + x_2) + y_1(x_2 - x_0) - y_2(x_0 + x_1)}{x_2 - x_0} \\ \mu &= y_0 - ax_0 - b \\ &= \frac{1}{2}\,\frac{y_0(x_2 - x_1) - y_1(x_2 - x_0) + y_2(x_1 - x_0)}{x_2 - x_0} \end{aligned}$$

If this procedure is used to find the best linear approximation to the data $(-1, 1)$, $(0, 0)$, and $(1, 1)$ then $a = 0$, $b = 1/2$ and $b_1(x) = 1/2$. Compare this with the best linear approximation to $f(x) = |x|$, $x \in [-1, 1]$ in Example 4.5.3. What happens in the present case if the data is changed to the set $(-1, 1)$, $(0, 0)$, $(1/2, 1/2)$? (Answer: $b_1(x) = (1-x)/3$.)

Example 4.5.7 Construct the best linear approximation to the data

$$S_4 = \{(-1, 1/2), (-1/2, 4/5), (1/2, 4/5), (1, 1/2)\}\,.$$

This reflects the situation in the construction preceeding Theorem 4.5.5; e.g. $m = 4$ and $n = 1$ (linear). There are $\binom{4}{3} = 4$ sets $S_{4,k}$, namely

$$\begin{aligned} S_{4,1} &= \{(-1, 1/2), (-1/2, 4/5), (1/2, 4/5)\} \\ S_{4,2} &= \{(-1, 1/2), (-1/2, 4/5), (1, 1/2)\} \\ S_{4,3} &= \{(-1, 1/2), (1/2, 4/5), (1, 1/2)\} \\ S_{4,4} &= \{(1/2, 4/5), (1/2, 4/5), (1, 1/2)\} \end{aligned} \tag{4.5.10}$$

As in the lines preceeding Theorem 4.5.5 label the points of $S_{4,1}$ $(x_0^1, y_0^1) = (-1, 1/2)$; $(x_1^1, y_1^1) = (-1/2, 4/5)$ and $(x_2^1, y_2^1) = (1/2, 4/5)$. Using (4.5.6) and (4.5.8) with $n = 1$ one finds

$$\hat{p}_2(x) = 4x^2 + 2x - 1 \quad \text{and} \quad p_2(x) = -2/5\,x^2 + 9/10\,.$$

Instead of using (4.5.6) to find $\hat{p}_2$ it is perhaps simpler to put $\hat{p}_2(x) = ax^2 + bx + c$ and require $\hat{p}_2(x_i^1) = (-1)^i$, $i = 0, 1, 2$. Similarly for the construction of $p_2(x)$. Hence, using (4.5.5) we have

$$\begin{aligned} p_{1,1}^*(x) &= p_2(x) - \mu_1 \hat{p}_2(x) \\ &= (-2/5 - 4\mu_1)x^2 - 2\mu_1 x + (9/10 + \mu_1) \end{aligned}$$

so that (requiring $p_{1,1}^* \in \Pi_1$) gives

$$\mu_1 = \frac{-2/5}{4} = -1/10$$

and

$$p_{1,1}^*(x) = 1/5\,x + 4/5\,.$$

In an exactly similar fashion one finds for the set $S_{4,2}$ in (4.5.10)

$$\hat{p}_2(x) = 8/3\,x^2 - 5/3 \quad \text{and} \quad p_2(x) = -2/5\,x^2 + 9/10$$

so that

$$\mu_2 = \frac{-2/5}{8/3} = -3/20$$

and

$$p_{1,2}^*(x) = 13/20\,.$$

You are asked in Problem **??** to conclude that $p_{1,3}^*$ and $p_{1,4}^*$ (best approximation on the sets $S_{1,3}$ and $S_{1,4}$, respectively) are given by $p_{1,2}^*$ and $-1/5\,x + 4/5$. Theorem 4.5.5 guarantees that $p_{1,2}^*(x)$ is the unique best approximation to the data S_4 ($|\mu_1| < |\mu_2|$).

As indicated by the previous examples computing the best approximate to a continuous function or the best approximate on a discrete point set can be an arduous process. To highlight the next theorem let us compute the Lagrange polynomial of degree two that interpolates $f(x) = |x|$ at the nodes $\{-\sqrt{3}/2, 0, \sqrt{3}/2\}$. The result is $L_2(x) = \frac{2\sqrt{3}}{3}\,x^2$. Let us compute

$$\begin{aligned} \delta_1(|x|) &= \max_{-1 \le x \le 1} \big|\,|x| - L_2(x)\big| \\ &= \max_{-1 \le x \le 1} \big|\,|x| - \frac{2\sqrt{3}}{3}\,x^2\big|\,. \end{aligned}$$

By considering the function $g(x) = x - \frac{2\sqrt{3}}{3}\,x$ on $[0, 1]$ and some elementary calculus one finds that

$$\max_{-1 \le x \le 1} |g(x)| = g\left(\frac{3}{4\sqrt{3}}\right) = \frac{\sqrt{3}}{8}$$

Figure 4.5.3: Best Linear Approximation to S_4.

so that (by symmetry)

$$\delta_1(|x|) = \frac{\sqrt{3}}{8} \cong .2165\,.$$

Recall, from Example 4.5.3, that $E_1(|x|) = 1/8 = .125$ (error in the best linear approximate to $|x|$). Hence the error in the Lagrange interpolant (on the "special" points of this example) is less than twice the error of the best approximate. The important point here is that the Lagrange interpolant is much easier to compute than is the best approximate. That this situation persists for larger degrees is the content of the following theorem.

Theorem 4.5.8 *Let $f \in C[-1,1]$ and put*

$$E_n(f) = \max_{-1\le x\le 1} |f(x) - b_n(x)| \le \max_{-1\le x\le 1} |f(x) - p_n(x)|$$

for all $p_n \in \Pi_n$. Denote by

$$x_j = \cos\left(\frac{(2j-1)\pi}{2n+2}\right) \quad , \quad j = 0, 1, \ldots, n+1$$

the zeros of $T_{n+1}(x) = \cos((n+1)\cos^{-1}(x))$ (see ??) and $L_n(x)$ the Lagrange polynomial satisfying

$$L_n(x_j) = f(x_j) \quad , \quad j = 1, \ldots, n+1\,.$$

Finally, put

$$\delta_n(f) = \max_{-1\le x\le 1} |f(x) - L_n(x)| \,.$$

Then

$$\delta_n(f) \le 4\, E_n(f) \ , \quad 1 \le n \le 20$$

and

$$\delta_n(f) \le 5\, E_n(f) \ , \quad 21 \le n \le 100 \,.$$

Example 4.5.9 If $f(x)$ is continuous and twice differentiable in $[a, b]$, and if $f''(x) > 0$ on $[a, b]$ then we can find the polynomial in b_n that satisfies equation (4.5.3) with $n = 1$, for all $b_n(x)$ in Π_n. Setting $b_1(x) = a_0 x + a_1$ where

$$\begin{aligned} a_0 &= \frac{f(b) - f(a)}{b - a} \\ a_1 &= \frac{f(c) - f(a)}{2} - \frac{c + a}{2}\, a_0 \end{aligned}$$

where c is determined such that $f'(a) = a_0$. Writing

$$r(x) = f(x) - a_0 x - a_1$$

we have

$$r'(x) = f'(x) - a_0$$

and $r'(x)$ is strictly increasing, since $f''(x) > 0$. There is a unique point c in $[a, b]$ such that $r'(c) = 0$. Thus $r(a) = -r(c) = r(b)$, i.e. b_1 is the best linear approximate of f.

Example 4.5.10 Symmetry can sometimes be used as an aid in constructing polynomials of best approximation. Consider for example, the approximation of $\cosh x$ in $[-1, 1]$ by a polynomial of degree ≤ 1.

¿From Figure 4.5.4 the best approximant of degree ≤ 1 is a polynomial of degree zero, i.e., the constant

$$c = \frac{1 + \cosh 1}{2} \cong 1.22$$

For, the deviations at $x = -1, 0, 1$ must be equal and alternating, i.e.

$$\cosh 1 - c = -[\cosh 0 - c] \quad (\cong .38)$$

which yields the value c given above. Notice that this result is a special case of Example 4.5.9.

The polynomial of best approximation $b_2(x)$ in Π_2 can be obtained in a similar way. By symmetry, the polynomial is even, and may be denoted by $ax^2 + b$. The difference, $r(x) = \cosh x - ax^2 - b$ in the half interval $[0, 1]$ has

Figure 4.5.4: The Approximation of $\cosh x$ by a Constant

maxima or minima where $r'(x) = \sinh x - 2ax = 0$, i.e. at $x = 0$, and at x_1, where $x_1 = \sinh x_1/2a$. The greatest positive deviations in $[0,1]$ are at $x = 0, 1$, the deviation at x_1 being negative. The equations

$$r(0) = -r(x_1) = r(1)$$

yield

$$1 - b = -[\cosh x_1 - a x_1^2 - b] = \cosh 1 - a - b$$

or

$$a \cong 54 \ , \ b \cong .995 \ , \ x_1 \cong .70$$

the greatest deviation is $1 - b \cong .005$.

It can be seen that the polynomial of degree ≤ 2 yields five extrema; the theory requires at least four. The reason for this is symmetry. The polynomial of best approximation of degree ≤ 3 is the same as the polynomial of best approximation of degree 2.

The approximation $b_0(x) = 1.22$ and $b_2(x) = .54x^2 + .995$ may be compared with the Taylor polynomial of degrees zero and 2, $q_0(x) = 1$, $q_2(x) = 1 + 1/2\, x^2$. For the latter polynomials we find that

$$\begin{aligned} \cosh x - 1 &\cong .60 \text{ at } x = 1 \\ \cosh x - 1 - \frac{1}{2}x^2 &\cong .04 \text{ at } x = 1\,. \end{aligned}$$

Ex. 5: The best straight line approximation of $\sinh x$ in $[-1, 1]$ is given by a line $b_1(x) = mx$ (the approximation must be odd in x). The difference,

$r(x) = \sinh x - mx$ has a minimum (Figure ???5.7) at x_1, where $r'(x_1) = \cosh x_1 - m = 0$. Consider the half–interval $[0, 1]$. Here the deviations at x_1 and 1 must be equal and opposite

$$\sinh 1 - m = -[\sinh x_1 - mx_1]$$

or

$$(1 + x_1) \cosh x_1 = \sinh 1 - \sinh x_1 \,..$$

Solving this equation, we have $x_1 = 0.49$ and $m = 1.12$. Here there are 4 extrema, at $\pm x_1$ and ± 1, and in analogy with Ex. 4, the best degree 2 polynomial approximation is the same as the best straight line approximation.

Figure 4.5.5: Approximation of $\sinh x$ by a straight line

Ex. 6: If $f(x)$ is continuous in $[a, b]$ and

$$m = \min_{x \in [a,b]} f(x) \ , \quad M = \max_{x \in [a,b]} f(x)$$

then the polynomial $b_0(x)$ in Π_0 that satisfies Theorem 4.5.2, is given by $b(x) = (m + M)/2$ and so

$$\max_{x \in [a,b]} |f(x) - b_0(x)| = 1/2(M - m) \,.$$

Example 4.5.11 The functions $\cos(\pi x/2)$, $1 - x^2$, and $1 - x^4$ all have 1/2 for the best–approximant of degree zero (cf. Ex. 6) in $[-1, 1]$. The function $(1 - x^2)(1 + 2x^2)$ has 9/16 for approximant of degree zero in $[-1, 1]$. In general, $b_n(x)$ ($b_n(x)$ in Π_n) is the best polynomial approximation of degree $\leq n$ to the set of functions $f_i(x)$ continuous in $[a, b]$ if the set of remainders $r_i(x) = f_i(x) - b_n(x)$ have common extrema at $n + 2$ points of $[a, b]$.

Example 4.5.12 If $T_n(x) = \cos n\theta$, where $x = \cos\theta$ then x is the best straight line approximation to $T_3(x) + x$, $T_5(x) + x$, $T_7(x) + x$, $\ldots$.

Problems

4.5.1 $*$ Write a computer algorithm to find the polynomial $p_n \in \Pi_n$ that best approximates $m > n + 1$ points in a minimax sense.

4.5.2 Show that the minimax approximation of $q_4(x) \equiv (x^2 - 1)(x^2 - 4)$ in $[-1, 1]$ by a quadratic $q_2(x)$ has the property that $r_4(x) \equiv q_4(x) - q_2(x)$ has 5 extrema of equal magnitude and alternating sign (the preceding theory requires at least 4). Sketch $r_4(x)$.

4.5.3 Approximate $q_5(x) = x^5$ in $[-1, 1]$ by a linear function, using the procedure $q_5(x)t_5(x) - (5/4)t_3(x)$ to eliminate the terms in x^5 and x^3. Show that the error $E(x) = q_5(x) - \{q_5(x) - t_5(x) - 5/4\, t_3(x)\}$ is bounded by $1/16 + (5/4)(1/4) = 3/8$ on $[-1, 1]$, and that this error bound is sharp. This procedure is called telescoping. Verify that the maximum and minimum points of $E(x)$ are at $-2^{1/4}/4$, $2^{1/4}/2$ and $-2^{1/4}/2$, $2^{1/4}/4$ respectively.

4.5.4 Approximate x^5 in $[-1, 1]$ by the linear function $5a^4x$ in the minimax sense. Show that the minimax approximation is achieved for some solution a of the equation $(a + 1)^2(4a^3 - 3a^2 + 2a - 1) = 0$. Show that the cubic on the left–hand side of this equation has one real zero $a \cong 0.6055$, and that the modulus of the maximum error of the approximation is approximately 0.326. Show also that the error curve has four extrema.

Chapter 5

Numerical Integration in One Variable

Introduction

Numerical integration of a function of a single variable is a highly developed technique. As in using a vending machine, we can choose what we need by pressing the appropriate button; so too in numerical integration, one has a choice of several very efficient computer programs. One can indeed reduce the choice to the Gauss integration schemes that work very well for a wide range of smooth functions, and even integrate functions that have mild discontinuities within the interval of integration—at a somewhat greater cost.

Digital computers free us from the labor of tedious calculations; but judgement is still called for in the use of numerical integration procedures if we wish to compute economically. The following situations are typical:

(i) we may need only a rough approximation to the value of the integral;

(ii) the integral can be expressed in closed form, but this form is not convenient for computation; and

(iii) the problem that gave rise to the integral may be more conveniently solved by other means.

Illustrations of these situations are given below. In particular, it is desirable to emphasize the value of rough approximations that can be obtained quickly. These are particularly valuable to the beginning student who is intent on finding out why certain procedures work well, and when. Furthermore, rough approximations are often desirable before undertaking extensive and expensive calculations.

In this chapter, considerable space is given to elementary methods that are useful for quick approximations, and also to methods based on the Euler-Maclaurin formula—although this formula is rarely used to evaluate integrals, it helps us to understand the success or failure of different methods. In contrast, much of the detailed analysis of Gaussian methods is omitted, since it is already available in many texts, and since it offers very little towards the understanding of practical difficulties.

5.1 Typical Examples

This section consists of a number of simple examples that illustrate the need for the user to think, and to proceed with caution.

Example 5.1.1 If $n > 2$, the integral

$$I(n) = \int_0^1 \frac{dx}{(1+x^n)^{1/2}}$$

cannot be expressed as an elementary function of n. However, $I(n)$ can be approximated numerically by many different methods.

The integral

$$J(n) = \int_0^1 \frac{dx}{(1-x^n)^{1/2}} \ , \quad n = 3, 4, 5, \ldots$$

is more difficult since it has an integrable singularity at $x = 1$. Integration by parts makes the numerical approximation of $J(n)$ more tractible. Setting

$$F(x) = \frac{1-x^n}{1-x}$$

and noting that $F(0) = 1$, we get

$$\begin{aligned} J(n) &= \int_0^1 (1-x)^{-1/2} F(x)^{-1/2} dx \\ &= 2 - \int_0^1 (1-x)^{1/2} F(x)^{-3/2} F'(x) dx \, . \end{aligned}$$

Example 5.1.2 It is sometimes possible to express the integral in closed form, in which case numerical integration is pointless. For example, it can be shown that

$$I = \int_0^\infty e^{-ax^2} \cos bx \, dx = \frac{1}{2} \left(\frac{\pi}{a}\right)^{1/2} \exp\left(-\frac{b^2}{4a}\right) \ , \quad a > 0 \, .$$

This closed form expression can be very useful, since it illustrates the dependence of I on a and b, and can therefore lead to a useful numerical method of approximating

$$\int_0^y e^{-ax^2} \cos bx \, dx$$

when y is finite.

Example 5.1.3 The integral

$$S(x) = \int_0^\infty e^{-t^2} \sin xt \, dt \tag{5.1.1}$$

cannot be expressed as a finite combination of elementary functions. If we differentiate the integral with respect to x, and integrate by parts, we obtain the differential equation

$$\frac{dS}{dx} = \frac{1}{2} - \frac{1}{2} x \, S(x) \ , \quad S(0) = 0 \, , \tag{5.1.2}$$

whose solution is given as a power series in x:

$$S(x) = \sum_{k=0}^{\infty} \frac{(-1)^k k\,!}{2(2k+1)\,!} x^{2k+1} \ , \quad |x| < \infty \, . \tag{5.1.3}$$

Five terms of the series (5.1.3) yield an accuracy of two significant figures if $|x| \leq 2$. We can also show that

$$S(x) = \frac{1}{x} + O\left(\frac{1}{x^3}\right) \tag{5.1.4}$$

as $x \to \infty$, and this information may be sufficient to solve the problem. If more accuracy is required we can use a Gaussian integration method.

Example 5.1.4

$$I = \int_0^x \frac{tdt}{a^3 + t^3} = \frac{1}{6a} \ln \left[\frac{x^2 - ax + a^2}{(x+a)^2} \right] + \frac{1}{a\sqrt{3}} \text{arc} \left(\tan \left(\frac{x\sqrt{3}}{2a - x} \right) \right) \, . \tag{5.1.5}$$

If the integral is to be evaluated for a single value of x and a, then the expression on the right is the most suitable method of computation. If I is required for several values of x and a, then numerical integration is preferable. Setting $t = xs$, $a = xb$ in (5.1.5), we get

$$I = \frac{1}{x} \int_0^1 \frac{sds}{b^3 + s^3} = \frac{2}{x} \int_{-1}^1 \frac{(1+v)dv}{8b^3 + (1+v)^3} \, , \tag{5.1.6}$$

where s has further been replaced by $(1+v)/2$ on the right.

Example 5: The differential equation

$$\frac{dy}{dx} = \frac{2y}{1 - x^4} \tag{5.1.7}$$

may be explicitly solved by method of separation of variables, by writing it in the form

$$\int \frac{dy}{y} = 2 \int \frac{dx}{1-x^4} \, .$$

Upon decomposing $(1-x^4)^{-1}$ into partial fractions, we find that (5.1.7) has the solution

$$y = A \left(\frac{1+x}{1-x} \right)^{1/2} e^{u(x)} \tag{5.1.8}$$

where A is a constant, and $u(x) = \arctan x$. This expression for y, although somewhat cumbersome, can be readily evaluated. However, if we add x to the right of (5.1.7), we get

$$\frac{dy}{dx} = \frac{2y}{1-x^4} + x \, ; \tag{5.1.9}$$

the solution of (5.1.9) using standard differential equation methods is

$$y = \left(\frac{1+x}{1-x} \right)^{1/2} e^{u(x)} \left[A + \int_0^x t \left(\frac{1-t}{1+t} \right)^{1/2} e^{-u(t)} dt \right] . \tag{5.1.10}$$

The integral on the right cannot be expressed in closed form, and it is moreover inconvenient for numerical evaluation. It is preferable, for solving (5.1.9), to use a numerical method of solving differential equations on (5.1.9) directly rather than to use a numerical method of approximating the integral in (5.1.10). When faced with a very cumbersome integral, it is often advisable to ask where it came from.

Example 5.1.5 The integral

$$J_n(x) = \frac{1}{\pi} \int_0^\pi \cos(x \cos\theta - n\theta) d\theta \tag{5.1.11}$$

can be easily evaluated numerically for moderate values of x and n, although numerical integration may be less efficient than the evaluation of $J_n(x)$ by use of the series

$$J_n(x) = \left(\frac{x}{2} \right)^n \sum_{k=0}^{\infty} \frac{(-x^2/4)^k}{k\,!\,(n+k)\,!} \, , \quad |x| < \infty \, . \tag{5.1.12}$$

If x or n (or both) are large, then the integrand in (5.1.11) oscillates very rapidly, and it is then preferable to use the asymptotic methods of Chapter 6.

Example 5.1.6 The integrals

$$\int_0^\infty \cos(ax^2)dx = \frac{1}{2}\left(\frac{\pi}{2a}\right)^{1/2} \quad , \quad \int_0^\infty \sin(ax^2)dx = \frac{1}{2}\left(\frac{\pi}{2a}\right)^{1/2} \quad , \tag{5.1.13}$$

converge because the integrand in each oscillates increasingly rapidly as x increases and the positive and negative contributions tend to cancel more and more for large x. The conditional convergence of each integral can be readily illustrated by means of the transformation $t = ax^2$. A numerical integration procedure that does not take account of the oscillation phenomenon of the integrands may yield poor results when applied to the numerical approximation of integrals of a similar type to (5.1.13).

Problems

5.1.1 Use the power series expansion of $S(x)$ in (5.1.1) and integrate termwise to get (5.1.3).

5.1.2 Either use integration by parts along with (5.1.1) or substituion along with (5.1.2) to show that

$$S(x) \sim \sum_{n=0}^{\infty} \frac{(-1)^n n\,!}{x^{2n+1}}$$

5.1.3 How many terms of the series (5.1.3) are required to evaluate to an accuracy of three significant figures for $|x| \le 10$?

5.1.4 Prove that the function

$$A(x) = \int_0^\infty e^{-t^3 - xt}\,dt$$

can be computed for intermediate and small values of x by means of the power series

$$A(x) = \frac{1}{3}\sum_{n=0}^{\infty} \frac{\Gamma\left(\frac{n+1}{3}\right)}{n\,!}(-1)^n x^n$$

while for large values of x, the asymptotic expansion

$$A(x) \sim \sum_{n=0}^{\infty} \frac{(-1)^n (3n)\,!}{n\,!\,x^{3n+1}}$$

is useful.

5.1.5 Prove that the function $f(x)$ defined by the integral

$$f(x) = \int_0^\infty \frac{e^{-u^2}}{x+u}\,du \quad , \quad x > 0$$

satisfies the differential equation $f' - 2xf = \pi^{1/2} - \dfrac{1}{x}$. Use this equation to deduce that

$$f(x) \sim \frac{1}{2}\sum_{n=0}^{\infty} \frac{(-1)^n \Gamma\left(\frac{n+1}{2}\right)}{x^{n+1}}$$

and also, that the function $y = f(x) + e^{-x^2}\log x$ satisfies the differential equation

$$y' + 2xy = \pi^{1/2} - \frac{1-e^{-x^2}}{x} = \pi^{1/2} + \sum_{n=1}^{\infty} \frac{(-1)^n x^{2n-1}}{n\,!}\,,$$

whose solution is $y = \sum\limits_{n=0}^{\infty} a_n x^n$ with $a_0 = -\frac{1}{2}\tau$, where τ is Euler's constant, i.e.

$$\tau = \lim_{k\to\infty}\left[\sum_{n=1}^{k} \frac{1}{n} - \ln(k)\right] = .577215665\,.$$

5.1.6 Use the identity

$$I = \int_0^\infty J_0(t)e^{-pt}dt = (1+p^2)^{-1/2}$$

which is obtainable from a table of Laplace transforms, to deduce that

$$\int_0^\infty J_0(t)dt = 1\,.$$

(Here J_0 is the Bessel function of order zero.) Can the transform

$$\int_0^\infty e^{-pt}\sin(t)dt = \frac{1}{1+p^2}$$

be used similarly?

5.2 The Equi-Spaced Midordinate and Trapezoidal Rules

The methods of this section are very simple, and very general. They are mainly used to obtain rough approximations, although they can be modified to yield more accurate approximations.

If $f(x)$ is continuous on the interval $[0,1]$, and if $f''(x)$ exists on the interval $(0,1)$, then it follows from Taylor's theorem that

$$f(x) = f\left(\frac{1}{2}\right) + \left(x - \frac{1}{2}\right) f'\left(\frac{1}{2}\right) + \frac{1}{2}\left(x - \frac{1}{2}\right)^2 f''(\zeta) \quad , \quad 0 \le x \le 1\,, \tag{5.2.1}$$

where ζ is between $1/2$ and x. Integrating both sides of (5.2.1), we get

$$\begin{aligned}\int_0^1 f(x)dx &= f\left(\frac{1}{2}\right) + \frac{1}{2}\int_0^1 \left(x - \frac{1}{2}\right)^2 f''(\zeta)dx \\ &= f\left(\frac{1}{2}\right) + \frac{1}{2} f''(\xi) \int_0^1 \left(x - \frac{1}{2}\right)^2 dx \\ &= f\left(\frac{1}{2}\right) + \frac{1}{24} f''(\xi)\,.\end{aligned} \tag{5.2.2}$$

In the second step of (5.2.2) we have used the mean value theorem for integrals, so that ξ is between 0 and 1. This yields the simplest form of the midordinate rule.

Example 5.2.1

$$I = \int_0^1 e^{-1/2\,x^2} dx = e^{-1/8} - \frac{1}{24}(1-\xi^2)e^{-1/2\,\xi^2}$$

for some $\xi \in (0,1)$. Notice that the error of approximating the integral by $e^{-1/8}$ is negative and bounded by $1/24 \cong .0416$. This bound is quite close to the exact error, since

$$I \cong .8557\ldots\ , \quad e^{-1/8} = .8825\ldots\ ,$$

so that the error, $I - e^{-1/8} \cong -.0262$.

The trapezoidal formula is obtained by a slightly different device. By means of Lagrange's interpolation formula we get

$$f(x) = (1-x)f(0) + xf(1) - \frac{x(1-x)}{2} f''(\zeta)\ , \quad 0 \le x \le 1\,. \tag{5.2.3}$$

Integrating both sides of (5.2.3), and using the mean value theorem, we obtain

$$\int_0^1 f(x)dx = \frac{1}{2}[f(0)+f(1)] - \frac{1}{12}f''(\xi) \tag{5.2.4}$$

for some $\xi \in (0,1)$. This is the simplest form of the trapezoidal rule.

Note that the error of (5.2.2) is $(1/24)f''(\xi_1)$ while the error of (5.2.4) is $-(1/12)f''(\xi_2)$ (in general, $\xi_1 \neq \xi_2$). *If $f''(x)$ has the same sign throughout $(0,1)$, then the errors of (5.2.2) and (5.2.4) are of opposite sign, and the magnitude of the difference bounds the error of each approximation,* i.e.

$$\begin{aligned} \left|\int_0^1 f(x)dx - \frac{1}{2}[f(0)+f(1)]\right| &\leq \left|f\left(\frac{1}{2}\right) - \frac{1}{2}[f(0)+f(1)]\right| \\ \left|\int_0^1 f(x)dx - f\left(\frac{1}{2}\right)\right| &\leq \left|f\left(\frac{1}{2}\right) - \frac{1}{2}[f(0)+f(1)]\right|. \end{aligned} \tag{5.2.5}$$

Example 5.2.2 Approximation of the integral of Example 5.2.1 by the trapezoidal rule (5.2.4) yields

$$\int_0^1 e^{-1/2\,x^2}dx = \frac{1}{2}(1+e^{-1/2}) + \frac{1}{12}(1-\xi^2)e^{-1/2\,\xi^2}$$

for some $\xi \in (0,1)$. The error of approximating the integral by the first term on the right is bounded by $1/12 \cong .0833$; the exact error is .0524.

The average of the midordinate and trapezoidal results is $1/2[.8825+.8023] \cong .8424$, and the error of this average is .0103. It may be shown that this error is less than the two preceding errors, since the second derivative of the integrand has the same sign throughout the interval.

Problems

5.2.1 Show that

(a) $\frac{1}{2} \leq \int_0^1 (1-x^2)^{1/2}dx \leq \left(\frac{3}{4}\right)^{1/2}$

(b) $\frac{1}{2}[\tan^{-1}(1)+\tan^{-1}(2)] \leq \int_0^1 \tan^{-1}(1+x)dx \leq \tan^{-1}\left(\frac{3}{2}\right)$.

5.2.2 Verify the statements in the last paragraph of the chapter for the integral

$$I = \int_0^1 \ln(x+1)dx = 2\,\ln\,2 - 1\,.$$

That is, approximate I via the midordinate and trapezoid rules, respectively.

5.3 Transformation of the Integral: Composite Rules

Let us consider the evaluation of

$$I = \int_a^b f(y)dy \tag{5.3.1}$$

where a and b are finite, $f(y)$ is continuous on $[a,b]$, and $f''(y)$ exists on (a,b). By means of the transformation

$$y = a + (b-a)x\,, \tag{5.3.2}$$

(5.3.1) becomes

$$I = (b-a)\int_0^1 f[a+(b-a)x]dx\,. \tag{5.3.3}$$

Noticing that

$$\frac{d^2}{dx^2}\,f[a+(b-a)x] = (b-a)^2\frac{d^2f(y)}{dy^2}$$

and using the notation

$$f''(y) = \frac{d^2}{dy^2}\,,$$

we can deduce from (5.2.2) and (5.2.4) that

$$\begin{aligned}\int_a^b f(y)dy &= (b-a)f\left(\frac{a+b}{2}\right) + \frac{(b-a)^3}{24}\,f''(\zeta)\\ &= \frac{(b-a)}{2}[f(a)+f(b)] - \frac{(b-a)^3}{12}\,f''(\xi)\end{aligned} \tag{5.3.4}$$

where ζ and ξ are points located on (a,b).

We can now obtain a more accurate approximation of $\int_0^1 f(x)dx$, by dividing the interval $[0,1]$ into n subintervals and applying one of (5.3.4) over each subinterval. Let j and n be positive integers such that $j \le n$. Then (5.3.4) yields

$$\begin{aligned}\int_{(j-1)/n}^{j/n} f(x)dx &= \frac{1}{n}\,f\left(\frac{2j-1}{2n}\right) + \frac{1}{24n^3}\,f''(\zeta_j)\\ &= \frac{1}{2n}\left[f\left(\frac{j-1}{n}\right) + f\left(\frac{j}{n}\right)\right] - \frac{1}{12n^3}\,f''(\xi_j)\end{aligned} \tag{5.3.5}$$

where ζ_j and ξ_j are located in $((j-1)/n, j/n)$.

Summing both sides of (5.3.5) over j, from 1 to n, we obtain

$$\begin{aligned}\int_0^1 f(x)dx &= \sum_{j=1}^{n}\int_{(j-1)/n}^{j/n} f(x)dx \\ &= M_n(f)+\frac{\epsilon_n}{24n^2} \quad \text{(midordinate)} \\ &= T_n(f)-\frac{\delta_n}{12n^2} \quad \text{(trapezoidal)}\end{aligned} \tag{5.3.6}$$

where

$$\begin{aligned}M_n(f) &= \frac{1}{n}\left[f\left(\frac{1}{2n}\right)+f\left(\frac{3}{2n}\right)+\cdots+f\left(\frac{2n-1}{2n}\right)\right] \\ T_n(f) &= \frac{1}{n}\left[\frac{1}{2}f(0)+f\left(\frac{1}{n}\right)+\cdots+f\left(\frac{n-1}{n}\right)+\frac{1}{2}f(1)\right]\end{aligned} \tag{5.3.7}$$

and where

$$\begin{aligned}\epsilon_n &= \frac{1}{n}[f''(\zeta_1)+f''(\zeta_2)+\cdots+f''(\zeta_n)] \\ \delta_n &= \frac{1}{n}[f''(\xi_1)+f''(\xi_2)+\cdots+f''(\xi_n)]\,.\end{aligned} \tag{5.3.8}$$

We now observe that ϵ_n is the average of the n values of $f''(\zeta_j)$, $j = 1, 2, \ldots, n$; hence if $f''(x)$ exists on $(0,1)$ there exists a point ζ such that

$$\epsilon_n = f''(\zeta)\ , \quad \zeta \in (0,1)\,.$$

Similarly, there exists a point ξ such that

$$\delta_n = f''(\xi)\ , \quad \xi \in (0,1)\,.$$

We have thus proved

Theorem 5.3.1 *Let $f(x)$ be continuous on the interval $[0,1]$, and let $f''(x)$ exist on $[0,1]$. Then*

$$\begin{aligned}\int_0^1 f(x)dx &= M_n(f)+\frac{1}{24n^2}f''(\zeta) \\ &= T_n(f)-\frac{1}{12n^2}f''(\xi)\end{aligned} \tag{5.3.9}$$

where M_n and T_n are defined in (5.3.7), and ζ and ξ are points on $(0,1)$.

From this theorem, and from the definitions of M_n and T_n, we deduce, as in (5.2.5):

Corollary 5.3.2 *If $f(x)$ is continuous on $[0,1]$ and $f''(x)$ exists and does not change sign on $(0,1)$, then*

$$\begin{aligned} \left| \int_0^1 f(x)dx - M_n(f) \right| &\leq |T_n(f) - M_n(f)| \\ \left| \int_0^1 f(x)dx - T_n(f) \right| &\leq |T_n(f) - M_n(f)| \, . \end{aligned} \tag{5.3.10}$$

We note furthermore, from the definition of T_n and M_n, that

$$T_{2n} = \frac{1}{2}[T_n + M_n] \, . \tag{5.3.11}$$

Theorem 5.3.1 has a simple geometric interpretation. The condition $f''(x) > 0$ on $(0,1)$ implies that $f(x)$ is convex on the interval $[0,1]$. Since a convex function lies below its chord, it is clear that

$$\int_0^1 f(x)dx < \frac{1}{2}[f(0) + f(1)] \text{ , and } f\left(\frac{1}{2}\right) < \int_0^1 f(x)dx \, .$$

These inequalities are illustrated in Figure 5.3.1.

Figure 5.3.1: A convex function

If $f(x)$ decreases for $x \geq 0$, then

$$f(1)+f(2)+\cdots+f(n) \leq \int_0^n f(x)dx \leq f(0)+f(1)+\cdots+f(n-1)\,,$$

the inequality being strict if $f(x)$ is strictly decreasing.

Many integration rules exist that are based on the evaluation of $f(x)$ at equi-spaced points. Very well known among these are Simpson's rule and Weddle's rule. Simpson's rule takes the form

$$\int_a^b f(x)dx = \frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]-\frac{(b-a)^5}{2880}f^{(4)}(\zeta)\,. \quad (5.3.12)$$

Composite Simpson's rules require that the interval be divided into an even number of strips, since the rule (5.3.12) uses three equi-spaced points, or two strips of length $(b-a)/2$. Weddle's is defined by

$$\begin{aligned}\int_a^b f(x)dx \quad &= \quad \frac{3h}{10}[\{f(a)+f(b)\}+5\{f(a+h)+f(a+5h)] \\ &\quad +\{f(a+2h)+f(a+4h)\}+6f(a+3h) \qquad (5.3.13) \\ &\quad -\frac{(b-a)^6}{6,531,840}f^{(6)}(\zeta)\,, \quad a<\zeta<b\,, \quad h=\frac{b-a}{6}\,.\end{aligned}$$

It involves six intervals of equal length, and seven evaluation points.

The formulas (5.3.12) and (5.3.13) were quite popular a few years ago, because of their simplicity, and because of the small coefficient of the error term. Weddle's rule may also be used to check Simpson's rule of every third step. Despite these advantages, the above rules are used little, save in rough approximations.

Corollary 5.3.2 suggests the following algorithm for computing $\int_0^1 f(x)dx$, when $f''(x)$ is of one sign on the interval $(0,1)$.

Algorithm 5.3.3 Quadrature Based on T_n and M_n

ϵ given
Evaluate T_1
$n=1$
    Evaluate M_n
        If $|T_n-M_n|<\epsilon$, print T_n or M_n
        If $|T_n-M_n|\geq\epsilon$
    $T_{2n}=1/2(T_n+M_n)$
$n\to 2n$
Return

Problems

5.3.1 Approximate

(a) $\frac{\pi}{4} = \int_0^1 (1-x^2)^{1/2} dx$

(b) $\int_0^1 \log(1+x)dx$

by $T_n((1-x^2)^{1/2})$ and by $M_n((1-x^2)^{1/2})$, for $n = 1, 2, 4, 8, 16$.

5.3.2 For the function $f(x) = (1-x)^{1/2}$, find positive constants c_1 and c_2 such that

$$\frac{c_1}{n^{3/2}} \leq \int_0^1 f(x)dx - T_n(f) \leq \frac{c_2}{n^{3/2}}\,.$$

5.4 The Euler-Maclaurin Formula

The Euler-Maclaurin formula is used little in applications, and our present treatment of it is therefore brief. Our treatment of this formula helps us explain why the trapezoidal and midordinate rules are sometimes so remarkably accurate, and it helps us explain the Romberg algorithm in Section 5.5, as well as the formulas of Section 5.6.

If we apply Taylor's formula

$$\begin{aligned} F(x) &= F(a) + (x-a)F'(a) + \cdots + \frac{(x-a)^{n-1}}{(n-1)\,!}F^{(n-1)}(a) \\ &\quad + \frac{1}{(n-1)\,!}\int_a^x (x-t)^{n-1}F^{(n)}(t)dt \end{aligned}$$

to the function $F(x) = \int_0^x f(t)dt$ then

$$\int_0^x f(t)dt = xf(0) + \int_0^x (x-t)f(t)dt\,. \qquad (5.4.1)$$

Similarly, applying Taylor's formula to $G(u) = \int_1^u f(t)dt$, we get

$$\int_1^u f(t)dt = (u-1)f(1) + \int_1^u (u-t)f(t)dt\,. \qquad (5.4.2)$$

Now by taking $x = 1$ in (5.4.1), $u = 0$ in (5.4.2), and adding the results, we get

$$\int_0^1 f(t)dt = \frac{1}{2}[f(0)+f(1)] + \int_0^1 \left(\frac{1}{2} - t\right) f'(t)dt\,. \qquad (5.4.3)$$

Let us set

$$P_1(x) = \frac{1}{2} - x \ , \quad 0 \leq x < 1 \ , \quad P_1(1+x) = P_1(x) \ , \quad \forall x \, . \tag{5.4.4}$$

With the notation $[x] =$ largest integer $\leq x$ the function $P_1(x)$ is given by

$$P_1(x) = -x + [x] + \frac{1}{2} \ , \quad \forall x \, . \tag{5.4.5}$$

This is graphed in Figure 5.4.1. It can be shown that $P_1(x)$ has a Fourier

Figure 5.4.1: $P_1(x) = -x + [x] + 1/2$

series expansion, given by

$$P_1(x) = \sum_{n=1}^{\infty} \frac{2\sin(2\pi n x)}{2\pi n} \, . \tag{5.4.6}$$

Integrating this series, and requiring $P_2(0) = P_2(1)$, we obtain

$$P_2(x) = -\sum_{n=1}^{\infty} \frac{2\cos(2\pi n x)}{(2\pi n)^2} \, .$$

Again, integrating this result and requiring $P_3(0) = P_3(1) = 0$ leads to

$$P_3(x) = -\sum_{n=1}^{\infty} \frac{2\sin(2\pi n x)}{(2\pi n)^3} \, .$$

Repeating these steps and requiring $P_{2r}(0) = P_{2r}(1)$ and $P_{2r+1}(0) = P_{2r+1}(1) = 0$ leads to the identities

$$\begin{aligned} P_{2r}(x) &= (-1)^r \sum_{n=1}^{\infty} \frac{2\cos(2\pi nx)}{(2\pi n)^{2r}} \\ P_{2r+1}(x) &= (-1)^r \sum_{n=1}^{\infty} \frac{2\sin(2\pi nx)}{(2\pi n)^{2r+1}}\,, \end{aligned} \tag{5.4.7}$$

$r = 1, 2, 3, \ldots$. We also deduce the following three relationships at once; that is

$$\begin{aligned} P_r(1-x) &= (-1)^r P_r(x) \\ (-1)^r P_{2r}(0) &> |P_{2r}(x)|\ , \quad 0 < x < 1\,, \end{aligned}$$

and

$$\begin{aligned} P_{2r}(x) - P_{2r}(0) &= (-1)^{r-1} \sum_{n=1}^{\infty} \frac{2[1-\cos(2\pi nx)]}{(2\pi n)^{2r}} \\ &= (-1)^{r-1} \sum_{n=1}^{\infty} \frac{4\sin^2(\pi nx)}{(2\pi n)^{2r}} \end{aligned} \tag{5.4.8}$$

Note that $P_{2r}(x) - P_{2r}(0) = 0$ for $x = 0$ and 1, and therefore $P_{2r}(x) - P_{2r}(0)$ assumes its largest modulus at $x = 1/2$. Further:

(i) It may be shown, using (5.4.7), that $P_{2r}(x)$ is an even function of $x - 1/2$ and $P_{2r+1}(x)$ an odd function of $x - 1/2$, and so $P_{2r+1}(1/2) = 0$.

(ii) Using (5.4.7) we can show that

$$P_{2r}\left(\frac{1}{2}\right) - P_{2r}(0) = -(1 - 2^{1-2r})P_{2r}(0)\,. \tag{5.4.9}$$

(iii) The numbers $P_{2r}(0)$ alternate in sign.

We remark that the functions $P_r(x)$ are related to the Bernoulli polynomials $B_r(x)$ (c.f. the problems at the end of this section) by the equation

$$-\,P_r(x) = \frac{B_r(x)}{r!}\ , \quad 0 \le x \le 1\,. \tag{5.4.10}$$

It is moreover useful to note that

$$- P_{2r}(1) = -P_{2r}(0) = (-1)^{r+1} \sum_{n=1}^{\infty} \frac{2}{(2\pi n)^{2r}} = \frac{B_{2r}(0)}{(2r)\,!} = \frac{B_{2r}}{(2r)\,!} \qquad (5.4.11)$$

where B_{2r} is the $2r^{\text{th}}$ Bernoulli number.

Integrating the extreme right of (5.4.3) by parts and using (5.4.5), we obtain:

Theorem 5.4.1 *Let $f^{(2m)}(x)$ be continuous on the interval $[0,1]$. Then*

$$\begin{aligned} \int_0^1 f(x)dx &= \frac{1}{2}[f(0)+f(1)] + \sum_{k=1}^{m} \frac{B_{2k}}{(2k)\,!}[f^{(2k-1)}(0) - f^{(2k+1)}(1)] \\ &\qquad + \int_0^1 \frac{B_{2m}(x)}{(2m)\,!}\, f^{(2m)}(x)dx\,. \end{aligned} \qquad (5.4.12)$$

We next use Theorem 5.4.1 and the transformation method of Section ????2.3 to obtain the formula

$$\begin{aligned} \int_{j-1/n}^{j/n} f(x)dx &= \frac{1}{2n}\left[f\left(\frac{j-1}{n}\right) + f\left(\frac{j}{n}\right)\right] \\ &\quad + \sum_{k=1}^{m} \frac{B_{2k}}{n^{2k}(2k)\,!}\left[f^{(2k-1)}\left(\frac{j-1}{n}\right) - f^{(2k-1)}\left(\frac{j}{n}\right)\right] \\ &\quad + \int_{j-1/n}^{j/n} \frac{B_{2m}(nx-j+1)}{n^{2m+1}}\, f^{(2m)}(x)dx\,. \end{aligned} \qquad (5.4.13)$$

Upon summing (5.4.13) with respect to j from 1 to n, we arrive at:

Theorem 5.4.2 *Let $f^{(2m)}(x)$ be continuous on the interval $[0,1]$. Then*

$$\begin{aligned} \int_0^1 f(x)dx &= T_n(f) + \sum_{k=1}^{m} \frac{B_{2k}}{n^{2k}((2k)\,!}[f^{(2k)}(0) - f^{(2k)}(1)] \\ &\qquad + \frac{1}{n^{2m+1}} \int_0^1 P_{2m}(nt) f^{(2m)}(t)dt,\,. \end{aligned} \qquad (5.4.14)$$

where $T_n(f)$ is defined in (5.3.7).

Proof: See Problem 5.4.5.

The coefficients $|B_{2k}/(2k)\,!|$ in the correction term decrease rapidly when k is large; from (5.4.7) we see that

$$\left|\frac{B_{2k}}{(2k)\,!}\right| = |P_{2k}(0)| \cong \frac{2}{(2\pi)^{2k}},\,. \tag{5.4.15}$$

Moreover $P_{2k+2}(0)/P_{2k}(0) \cong 1/40$. The first five coefficients are given in Table 5.4.1.

Table 5.4.1: The Coefficients $P_{2k}(0)$.

k	1	2	3	4	5
$-P_{2k}(0) = \frac{B_{2k}}{(2k)!}$	$\frac{1}{12}$	$-\frac{1}{720}$	$\frac{1}{30{,}240}$	$-\frac{1}{1{,}209{,}600}$	$\frac{1}{47{,}900{,}160}$

We note in particular, that the right-hand side of (5.4.14) is an asymptotic expansion of $\int_0^1 f(x)dx$ as $n \to \infty$. For the first few terms of this Table 5.4.1 records

$$\begin{aligned}\int_0^1 f(x)dx \;\; = \;\; & T_n(f) - \frac{1}{12n^2}\,[f'(1) - f'(0)] \\ & + \frac{1}{240n^4}\,[f^{(3)}(1) - f^{(3)}(0)] - \frac{1}{30,240}[f^{(5)}(1) - f^5(0)] \\ & + \frac{1}{1,209,600n^8}\,[f^{(7)}(1) - f^{(7)}(0)] \\ & - \frac{1}{47,900,160n^{10}}\,[f^{(9)}(1) - f^{(9)}(0)]\end{aligned} \tag{5.4.16}$$

as $n \to \infty$.

Example 5.4.3 We use the formula (5.4.16) to obtain an approximation of the integral

$$\begin{aligned} J_0(2) & = \frac{1}{\pi}\int_0^\pi \cos[2\sin\theta]d\theta \\ & = \int_0^1 \cos[2\sin\pi\, x]dx,\,. \end{aligned} \tag{5.4.17}$$

In this case, we easily prove by induction that $(\cos[2\sin\pi x])^{(2k-1)} = 0$ for $x = 0$ and $x = 1$, and for every positive integer k. At first glance, (5.4.16) leads us to suspect that the integral (5.4.17) is equal to its trapezoidal sum, for $n = 1, 2, 3, \ldots$. This is not the case. Equation (5.4.14) merely tells us that as $n \to \infty$, the error, $J_0(x) - T_n(\cos[2\sin\pi x])$ approaches zero faster than any power of $1/n$. The results of the computations are given in Table 5.4.2.

Table 5.4.2: The approximation of $J_0(2)$.

n	$T_n(\cos[\sin\pi x])$	$J_0(2)$
2	.29192	
3	.22630	.22389079
4	.223935	
6	.22389079	
8	.22389078	
12	.22389078	

The Euler-Maclaurin formula (5.4.14) is sometimes written in the form

$$\int_0^{nh} f(t)dt = T_h(f) - C(2m, nh) + R(2m, nh) \tag{5.4.18}$$

where

$$\begin{aligned} T_h(f) &= h\left[\frac{1}{2}f(0) + f(h) + \cdots + f((n-1)h) + \frac{1}{2}f(nh)\right] \\ C(2m, nh) &= \sum_{k=1}^{m} h^{2k} P_{2k}(0)[f^{(2k-1)}(nh) - f^{(2k-1)}(0)] \qquad (5.4.19) \\ R(2m, nh) &= h^{2m+1}\int_0^{nh} P_{2m}(t/h) f^{(2m)}(t)dt\,, \end{aligned}$$

this formula being valid, for example, if $f \in C^{(2m)}[0, nh]$.

If we let $n \to \infty$ in (5.4.18), we arrive at:

Theorem 5.4.4 *Let $\int_0^\infty |f(t)|dt$ exist, and let $\int_0^\infty |f^{(2m)}(t)|dt < \infty$. Then*

$$\int_0^\infty f(t)dt = h\left[\frac{1}{2}f(0) + f(h) + f(2h) + \cdots\right]$$

$$
\begin{aligned}
&+\sum_{k=1}^{m} h^{2k} P_{2k}(0) f^{(2k-1)}(0) \qquad (5.4.20)\\
&+h^{2m+1}\int_0^\infty P_{2m}(t/h) f^{(2m)}(t)dt,.
\end{aligned}
$$

Corollary 5.4.5 *Let* $\int_0^\infty |f(t)|dt$ *exist, and let* $\int_0^\infty |f^{(2m)}(t)|dt < \infty$ *for* $m = 1, 2, 3, \ldots$. *Then*

$$
\begin{aligned}
\int_0^\infty f(t)dt \quad &\sim \quad h\left[\frac{1}{2} f(0) + f(h) + f(2h) + \cdots\right] \\
&+\sum_{k=1}^{m} h^{2k} P_{2k}(0) f^{2k-1}(0)
\end{aligned}
\qquad (5.4.21)
$$

as $h \to 0$.

Let us next replace $f(t)$ in Theorem 5.4.4 by $g(t+a)$. Letting $a \to -\infty$, we arrive at:

Theorem 5.4.6 *Let* $\int_{-\infty}^\infty |f(t)|dt$ *exist, and let* $\int_{-\infty}^\infty |f^{2m}(t)|dt < \infty$. *Then*

$$
\begin{aligned}
\int_{-\infty}^\infty f(t)dt \quad &= \quad h\sum_{k=-\infty}^{\infty} f(kh) \\
&+h^{2m+1}\int_{-\infty}^\infty P_{2m}\left(\frac{t}{h}\right) f^{(2m)}(t)dt\,.
\end{aligned}
\qquad (5.4.22)
$$

Corollary 5.4.7 *Let* $\int_{-\infty}^\infty |f(t)|dt$ *exist, and let* $\int_{-\infty}^\infty |f^{(2m)}(t)|dt < \infty$ *for* $m = 1, 2, 3, \ldots$. *Then*

$$
\int_{-\infty}^\infty f(t)dt - h\sum_{k=-\infty}^{\infty} f(kh) = 0(h^\alpha) \qquad (5.4.23)
$$

as $h \to 0$ *where* $\alpha > 0$ *is arbitrary.*

Example 2: The result of Corollary 5.4.5 may be used not only to approximate integrals, but also to evaluate some slowly convergent series. For example, to evaluate

$$
S = \sum_{n=0}^{\infty} (1/(n+1)^2)
$$

correct to 8 decimals we would need to sum about 10,000 terms of the series. If we set $h = 1$ in (5.4.21), we get

$$\begin{aligned} S = \sum_{n=0}^{\infty} \frac{1}{(1+n)^2} &\cong \int_0^{\infty} \frac{dx}{(1+x)^2} + \frac{1}{2} + \sum_{k=1}^{m} (-1)^{k-1} B_{2k} \\ &= \frac{3}{2} + \sum_{k=1}^{m} (-1)^{k-1} B_{2k} \end{aligned} \tag{5.4.24}$$

since $f^{2k-1}(0) = (-1)^k (2k)\,!$. The error between the true result $(S = \pi^2/6)$ and the approximation as a function of m on the extreme right of (5.4.24) is given in Table 5.4.3.

Table 5.4.3: The error of the approximation (??? 5.4.24)

m	0	1	2	3	4	5
Error	.145	-.022	.012	-.012	.021	-.055

Let us next approximate

$$S_1 = \sum_{n=0}^{\infty} 1/(10+n)^2 \,.$$

We obtain

$$S_1 = \sum_{n=0}^{\infty} \frac{1}{(10+n)^2} \cong \int_0^{\infty} \frac{dx}{(10+x)^2} + \frac{1}{2} \cdot \frac{1}{10^2} + \sum_{k=1}^{m} \frac{B_{2k}}{10^{2k+1}} \,. \tag{5.4.25}$$

With $m = 2$, the right of (5.4.25) yields

$$S_1 \cong \frac{1}{10} + \frac{1}{200} + \frac{1}{6} \cdot \frac{1}{10^3} - \frac{1}{30 \cdot 10^5} \cong \cdot 1051663333 \,.$$

We now add this result to the directly calculated sum

$$\sum_{n=1}^{9} \frac{1}{n^2} \cong 1.5397677311$$

to get

$$S = 1.6449340644$$

which is correct to 8 decimals.

This result illustrates very well the asymptotic nature of the Euler-Maclaurin formula. Although the series

$$\sigma = \sum_{k=1}^{\infty} \frac{B_{2k}}{a^{2k+1}}$$

diverges for every $a > 0$ when a is large, the series begins to converge very rapidly; for the case of (5.4.25), it may be shown that the error upon truncating σ after m terms is less in magnitude than the magnitude of the first neglected term.

Example 3:

$$I = \int_1^{n+1} \log t \, dt = \int_0^n \log(1+t) dt \, .$$

For this case,

$$\frac{d^{2k-1}}{dt^{2k-1}} (\log(t+1)) = \frac{(2k-1)\,!}{(t+1)^{2k-1}} \, , \quad k \geq 1 \, .$$

All of these derivatives have the same sign throughout the interval of integration; this fact enables us to conclude that if we use $h = 1$ in (5.4.18) then

$$\begin{aligned} I &= \log 2 + \log 3 + \cdots + \log n + \frac{1}{2} \log(n+1) \\ &\quad - \sum_{k=1}^{m} P_{2k}(0) \left[\frac{1}{(n+1)^{2k-1}} - 1 \right] (2k-1)\,! \, , \end{aligned}$$

the error is less than the first neglected term. Also, I can be explicitly evaluated to yield $I = (n+1)\log(n+1) - n$. Hence we obtain using (5.4.19)

$$\log((n+1)\,!) = \left(n + \frac{3}{2}\right) \cdot \log(n+1) - n - C(2m,n) + R(2m,n)$$

where

$$|R(2m,n)| < (2m-2)\,! \, |P_{2m}(0)| \left[1 - \frac{1}{n} 2m - 1 \right] .$$

Example 4: The two integrals

$$\int_0^\infty e^{-t} \cos xt \, dt \quad \text{and} \quad \int_0^\infty e^{-t} \sin xt \, dt$$

can be handled simultaneously by means of

$$I(x) = \int_0^\infty e^{-t(1-ix)} dt\,.$$

For the integrand in this integral,

$$\frac{d^k}{dt^k}\left(e^{-t(1-ix)}\right) = (-1)^k (1-ix)^k e^{-t(1-ix)}$$

and so

$$\left|\frac{d^k}{dt^k} e^{-t(1-ix)}\right| \le (1+x^2)^{k/2} e^{-t}\,.$$

Consequently, using (5.4.20),

$$\begin{aligned} I(x) &= h\left[\frac{1}{2} + e^{-h(1-ix)} + e^{-2h(1-ix)} + \cdots\right] \\ &\quad - \sum_{k=1}^{m} h^{2k} P_{2k}(0)(1-ix)^{2k-1} + R(2m,\infty) \end{aligned}$$

where

$$R(2m,\infty) = h^{2m+1} \int_0^\infty P_{2m}(t/h)(1-ix)^{2m} e^{-t(1-ix)} dt$$

so that

$$|R(2m,\infty)| \le h^{2m+1} |P_{2m}(0)| (1+x^2)^m\,.$$

For fixed m and x, $R(2m,\infty)$ can thus be chosen as small as we please, by taking h sufficiently small.

Example 5:

$$I = \int_{-\infty}^{\infty} e^{-t^2} dt$$

The results of Corollary 5.4.7 apply to this integral, and we may therefore expect the approximation

$$I \cong h \sum_{k=-\infty}^{\infty} e^{-k^2 h^2}$$

to be quite accurate. That this is indeed the case, is illustrated in Table 5.4.4.

Table 5.4.4: $\int_{-\infty}^{\infty} exp(-t^2)\, dt = 1.77245385$

h	$I \simeq h \sum_{-\infty}^{\infty} e^{k^2 h^2}$
1	1.77263720
1/2	1.77245377

For this problem we can get an estimate on the rate of convergence, using the identity

$$I(x) = \int_{-\infty}^{\infty} e^{-t^2} \cos xt\, dt = (2\pi)^{1/2} e^{-x^2/4}$$

which yields

$$\left(\frac{d}{dx}\right)^{2m} I(x) = \int_{-\infty}^{\infty} e^{-t^2} t^{2m} \cos(xt) dt = (2\pi)^{1/2} \frac{d^{2m}}{dx^{2m}} \left(e^{-x^2/4}\right) . \quad (5.4.26)$$

Thus, using (5.4.22), the error with $m = 1$ is

$$\epsilon = h^3 \int_{-\infty}^{\infty} P_2(t/h)[4t^2 - 2] e^{-t^2} dt\, .$$

Taking only the first term in the Fourier series of P_2 in (5.4.7), we get

$$\begin{aligned} \epsilon &\cong h \int_{-\infty}^{\infty} \frac{2\cos(2\pi t/h)}{4\pi^2} [4t^2 - 2] e^{-t^2} dt \\ &= \frac{2(2\pi)^{1/2}}{\pi^2} h(\pi^2 - h^2) e^{-\pi^2/h^2} \cong 5h\, e^{-\pi^2/h^2} , \end{aligned}$$

where we have used (5.4.26) to evaluate the integral. Hence $\epsilon \cong 2 \cdot 5 \times 10^{-4}$ with $h = 1$, while $\epsilon \cong 3 \times 10^{-17}$ with $h = 1/2$.

Problems

5.4.1 Derive the series (5.4.6).

5.4.2 Define the Bernoulli polynomials $B_n(x)$ by means of the generating function

$$\frac{ze^{xz}}{e^z - 1} = \sum_{n=0}^{\infty} \frac{B_n(x)}{n!} z^n .$$

Differentiate each side to conclude that

$$\frac{z^2 e^{xz}}{e^z - 1} = \sum_{n=1}^{\infty} B_n(x)\frac{z^n}{n!} = \sum_{n=1}^{\infty} n\, B_{n-1}(x)\frac{z^n}{n!},$$

i.e., that $B_n(x) = n\, B_{n-1}(x)$.

5.4.3 Define $B_n = B_n(0)$, and conclude that

$$B_n(x) = B_n + n\int_0^x B_{n-1}(t)dt\ ,\quad n = 1, 2, 3, \ldots .$$

5.4.4 Verify (5.4.10).

5.4.5 Derive the equation (5.4.14), using (5.4.13).

5.4.6 Use (5.4.21) to evaluate $\displaystyle\sum_{n=1}^{\infty} 1/n^3$ correct to eight decimals.

5.4.7 Evaluate $\displaystyle\int_0^{\infty} e^{-x^2}\cos 2x\, dx$ correct to five significant figures.

5.4.8 Derive the formula

$$\begin{aligned} h\int_0^1 f(a+\zeta h)d\zeta - hf\left(a+\frac{1}{2}h\right) &= \sum_{r=1}^{m} h^{2r}P_{2r}\left(\frac{1}{2}\right)[f^{(2r-1)}(a+h) - f^{(2r-1)}(a)] \\ &\quad + h^{2m+1}\int_0^{nh} P_{2m}\left(\zeta - \frac{1}{2}\right) f^{(2m)}(a+\zeta h)d\zeta \end{aligned}$$

which is valid, for example, if $f \in C^{(2m)}[a, a+h]$.

5.4.9 Use the results of Problem 5.4.8 to deduce the formula

$$\begin{aligned} h\int_0^{nh} f(a &+\zeta h)d\zeta - h\textstyle\sum_{k=1}^{n} f\left(a + \frac{2k-1}{2}h\right) \\ &= \textstyle\sum_{r=1}^{m} h^{2r}P_{2r}\left(\frac{1}{2}\right)[f^{(2r-1)}(a+nh) - f^{(2r-1)}(a)] \\ &\quad + h^{2m+1}\textstyle\int_0^{nh} P_{2m}\left(\zeta - \frac{1}{2}\right) f^{(2m)}(a+\zeta h)d\zeta\,. \end{aligned}$$

5.4.10 Use the results of Problem 5.4.9 to deduce that if $f \in C^{(m)}[0,1]$ for every $m = 1, 2, 3, \ldots$, then

$$\int_0^1 f(x)dx = M_n(f) + \frac{1}{24n^2}[f'(1) - f'(0)] - \frac{7}{5760\,n^4}[f^{(3)}(1) - f^{(3)}(0
+\frac{31}{967,680\,n^6}[f^{(5)}(1) - f^{(5)}(0)]$$

$$-\frac{127}{154,828,800\,n^8}[f^{(7)}(1) - f^{(7)}(0)] + 0(n^{-10})$$

as $n \to \infty$, where $M_n(f)$ is defined in (5.3.7).

5.4.11 Combine (5.4.26) and the result of Problem 5.4.8 that if

$$\int_0^\infty |f^{(2m)}(t)|dt < \infty$$

for $m = 1, 2, 3, \ldots$ then

$$\int_0^\infty f(t)dt - \sum_{n=1}^\infty f\left(\frac{2n-1}{2}h\right) = -\frac{h^2}{24}f'(0) + \frac{7h^4}{5760}f^{(3)}(0)$$
$$-\frac{31}{967,680}f^{(5)}(0) + \frac{127\,h^8}{154,828,800}f^{(7)}(0) + 0(h^{10})$$

as $h \to 0$.

5.4.12 $*$ Show that $B_k\left(\frac{1}{2}\right) = 2(2^{-k} - 1)B_k$.

5.5 The Romberg Algorithm

In practice it is possible to achieve only limited accuracy by the trapezoidal and midordinate formulas, because of the large number of points that are required and because of roundoff error. The Romberg method attempts to avoid this problem by producing a more rapidly convergent expansion of $\int_0^1 f(x)dx$. Starting from (5.4.16) we see that

$$\int_0^1 f(x)dx \sim T_n(f) + \frac{a_1}{n^2} + \frac{a_2}{n^4} + \frac{a_3}{n^6} + \cdots \tag{5.5.1}$$

where the a's depend on f but not on n. If we replace n by $2n$ in this relation, then

$$\int_0^1 f(x)dx \sim T_{2n}(f) + \frac{a_1}{4n^2} + \frac{a_2}{16n^4} + \frac{a_3}{64n^6} + \cdots . \tag{5.5.2}$$

We multiply (5.5.2) by 2^2, subtract (5.5.1), and divide the result by $2^2 - 1$, to get

$$\int_0^1 f(x)dx \sim \frac{2^2 T_{2n}(f) - T_n(f)}{2^2 - 1} + \frac{1}{2^2 - 1}\left[\left(\frac{2^2}{2^4} - 1\right)\frac{a_2}{n^4} + \cdots\right] . \tag{5.5.3}$$

If we use the notation $T_n^{(1)} = T_n(f)$, and

$$T_{2n}^{(2)} = \frac{2^2 T_{2n}^{(1)} - T_n^{(1)}}{2^2 - 1} \tag{5.5.4}$$

we see that, while

$$\int_0^1 f(x)dx = T_n^{(1)} + 0\left(\frac{1}{n^2}\right) \quad , \quad n \to \infty , \tag{5.5.5}$$

$$\int_0^1 f(x)dx = T_{2n}^{(2)} + 0\left(\frac{1}{n^4}\right) \quad , \quad n \to \infty . \tag{5.5.6}$$

In general, if we define $T_m^{(j+1)}$ by

$$T_{2^k n}^{(j+1)} = \frac{2^{2j} T_{2^k n}^{(j)} - T_{2^{k-1} n}^{(j)}}{2^{2j} - 1} \tag{5.5.7}$$

for $j = 1, 2, 3, \ldots$, and $k = j, j+1, j+2, \ldots$, then it may be shown that

$$\int_0^1 f(x)dx = T_{2^k n}^{(j+1)} + 0\left(\frac{1}{n^{2j+2}}\right) \quad , \quad n \to \infty . \tag{5.5.8}$$

The following table illustrates the method of tabulating the sequence of numbers.

It may be shown that in theory, each of the columns in Table 5.5.1 converges to $\int_0^1 f(x)dx$, the rate of convergence of the j'th column being $0(n^{-2j})$ as $n \to \infty$.

In order to get the j'th column in Table 5.5.1, we have of course assumed that $f^{(2j-1)}(x)$ exists on the interval $[0, 1]$. This does not hold for every integer j, for example, when $f(x) = x^a$, and a is not an integer. Also, the columns of the Romberg table may appear to converge initially—to the wrong value.

Table 5.5.1: The Romberg Table

$T_n^{(1)}$				
$T_{2n}^{(1)}$	$T_{2n}^{(2)}$			
$T_{2^2n}^{(1)}$	$T_{2^2n}^{(2)}$	$T_{2^2n}^{(3)}$		
$T_{2^3n}^{(1)}$	$T_{2^3n}^{(2)}$	$T_{2^3n}^{(3)}$	$T_{2^3n}^{(4)}$	
$T_{2^4n}^{(1)}$	$T_{2^4n}^{(2)}$	$T_{2^4n}^{(3)}$	$T_{2^4n}^{(4)}$	$T_{2^4n}^{(5)}$

Example 5.5.1 Let us evaluate

$$\log 3 = \int_0^1 \frac{2dx}{2x+1} \cong 1.098612$$

by use of the Romberg algorithm. The results we obtain are given in Table 5.5.2.

Table 5.5.2: The Romberg Algorithm

j	$T_{2^j}^{(1)}$	$T_{2^j}^{(2)}$	$T_{2^j}^{(3)}$	$T_{2^j}^{(4)}$
0	1.333333			
1	1.166667	1.111111		
2	1.116667	1.100000	1.099259	
3	1.103211	1.098725	1.098641	1.098631

Problems

5.5.1 Explain why the Romberg method fails for the case of

$$\int_0^1 x \log x dx \quad \text{or} \quad \int_0^1 (x^{1/3} + x^{1/4}) dx$$

5.5.2 How is $T_{2^k n}^{(2)}$ related to Simpson's rule (see(5.3.12))?

5.5.3 Derive the Romberg algorithm using the midordinate, rather than the trapezoid rule. (Solution: Simply replace the T's with M's in the equations of this section. Why?)

5.5.4 Apply the Romberg method to approximate

(a) $\int_0^1 x^{1/2} dx$

(b) $\int_0^1 \log(x) dx$

(For the case of (b), use $M_{2^k n}^{(j+1)}$ procedure of Problem 5.5.3 as well as the $T_{2^k n}^{(j+1)}$ procedure of the chapter.)

5.6 Gaussian Quadrature

Let $w(x) > 0$ on the interval (a, b), and let us introduce the inner product notation

$$(f, g) = \int_a^b w(x) f(x) g(x) dx \tag{5.6.1}$$

where f and g are continuous functions on (a, b). We assume that the numbers

$$\mu_k = (x^k, 1) = \int_a^b w(x) x^k dx \tag{5.6.2}$$

are finite for every integer $k \geq 0$. We can then use the Gram-Schmidt process to construct a sequence of polynomials $\{\psi_k\}_{k=0}^{\infty}$ that are orthonormal over (a, b) with respect to $w(x)$; that is,

$$(\psi_k, \psi_\ell) \equiv \int_a^b w(x) \psi_k(x) \psi_\ell(x) dx = \delta_{k\ell} = \begin{cases} 1 & k = \ell \\ 0 & k \neq \ell \end{cases}. \tag{5.6.3}$$

In particular, we may assume that

$$\psi_n(x) = k_n x^n + s_n x^{n-1} + \pi_{n-2}(x) \tag{5.6.4}$$

where $k_n > 0$, and where $\pi_{n-2}(x)$ is a polynomial of degree $n - 2$ in x.

The Gram-Schmidt construction is based on the same principles as in Chapter 2 and proceeds as follows: Define

$$\begin{aligned} \psi_0(x) &= (1, 1)^{-1/2} \\ \psi_n(x) &= (\theta_n, \theta_n)^{-1/2} \theta_n(x) \end{aligned} \tag{5.6.5}$$

where

$$\theta_n(x) = x^n - \sum_{j=0}^{n-1} (x^n, \psi_j)\psi_j(x) \tag{5.6.6}$$

then the $\{\psi_n\}_0^\infty$ satisfy (5.6.3) (Problem 5.5.1). It can be shown that the n zeros of $\psi_n(x)$ are real, distinct and lie within the interval (a, b) (Problem 5.5.3). This fact plays a role in the quadrature formula based on the $\{\psi_n\}_{n=0}^\infty$. If we multiply the Lagrange polynomial that interpolates a function f at the zeros x_j of ψ_n by $w(x)$ and integrate over (a, b), we obtain

$$\begin{aligned} \int_a^b w(x)f(x)dx &\cong \int_a^b w(x) \sum_{j=1}^n f(x_j) \left[\frac{\psi_n(x)}{(x - x_j)\psi_n'(x_j)}\right] dx \\ &= \sum_{j=1}^n w_j f(x_j) \end{aligned} \tag{5.6.7}$$

where

$$w_j = \int_a^b w(x) \frac{\psi_n(x)}{(x - x_j)\psi_n'(x_j)} dx\,. \tag{5.6.8}$$

The approximation (5.6.7) with w_j chosen as in (5.6.8) is known as Gaussian quadrature.

We now derive some properties of Gaussian quadrature.

Lemma 5.6.1 $w_j > 0$, $j = 1, 2, \ldots, n$.

Proof: We multiply the integrand in (5.6.7) by

$$1 = \sum_{k=1}^n \frac{\psi_n(x)}{(x - x_k)} \psi_n'(x_k) \tag{5.6.9}$$

to obtain

$$\begin{aligned} w_j &= \int_a^b \left(\sum_{k=1}^n \frac{\psi_n(x)}{(x - x_k)\psi_n'(x_k)}\right) \frac{\psi_n(x)}{(x - x_j)\psi_n'(x_j)} \\ &\equiv \left(\sum_{k=1}^n \frac{\psi_n(x)}{(x - x_k)\psi_n'(x_k)}\,,\quad \frac{\psi_n(x)}{(x - x_j)\psi_n'(x_j)}\right) \\ &= \left(\frac{\psi_n(x)}{(x - x_j)\psi_n'(x_j)}\,,\quad \frac{\psi_n(x)}{(x - x_j)\psi_n'(x_j)}\right) \end{aligned} \tag{5.6.10}$$

$$+ \sum_{\substack{k=1 \\ k \neq j}}^{n} \left(\psi_n(x) \ , \ \frac{\psi_n(x)}{(x - x_j)(x - x_k)\psi_n'(x_j)\psi_n'(x_k)} \right)$$

$$= \left(\frac{\psi_n(x)}{(x - x_j)\psi_n'(x_j)} \ , \ \frac{\psi_n(x)}{(x - x_j)\psi_n'(x_j)} \right)$$

since each of the polynomials

$$P_{jk}(x) = \frac{\psi_n(x)}{(x - x_j)(x - x_k)\psi'(x_j)\psi'(x_k)} \tag{5.6.11}$$

is of degree less than n, which implies (c.f. Problem 5.6.2) that $(\psi_n, P_{jk}) = 0$. On the other hand, the term on the extreme right of (5.6.10) is positive, and so $w_j > 0$.

Lemma 5.6.2 *Let $P_{2n-1}(x)$ be any polynomial of degree $2n - 1$ in x. Then there are polynomials $q_{n-1}(x)$ and $r_{n-1}(x)$, both of degree $n - j$ in x, such that*

$$P_{2n-1}(x) = \psi_n(x) q_{n-1}(x) + r_{n-1}(x) \, . \tag{5.6.12}$$

Proof: This is left as a problem (Problem 5.6.5).

Lemma 5.6.3 *The formula*

$$\int_a^b w(x) f(x) dx = \sum_{j=1}^{n} w_j f(x_j) \tag{5.6.13}$$

holds whenever $f(x)$ is any polynomial of degree $2n - 1$ in x.

Proof: If $r_{n-1}(x)$ is any polynomial of degree $n - 1$ in x, then (c.f. Chapter 2)

$$r_{n-1}(x) = \sum_{j=1}^{n} \frac{\psi_n(x) r_{n-1}(x_j)}{(x - x_j) r_{n-1}^1(x_j)} \, . \tag{5.6.14}$$

It follows therefore from (5.6.7) that

$$\int_a^b w(x) r_{n-1}(x) dx = \sum_{j=1}^{n} w_j r_{n-1}(x_j) \, . \tag{5.6.15}$$

On the other hand, if $P_{2n-1}(x)$ is an arbitrary polynomial of degree $2n-1$ in x, then using Lemma 5.6.2, we get

$$\begin{aligned}\int_a^b w(x)P_{2n-1}(x)dx &= \int_a^b w(x)\psi_n(x)q_{n-1}(x)dx \\ &\quad + \int_a^b w(x)r_{n-1}(x)dx \\ &= \int_a^b w(x)r_{n-1}(x)dx\end{aligned} \tag{5.6.16}$$

since $(\psi_n, q_{n-1}) = 0$. We similarly note that

$$\begin{aligned}\sum_{j=1}^n w_j P_{2n-1}(x_j) &= \textstyle\sum_{j=1}^n w_j[\psi_n(x_j)q_{n-1}(x_j) + r_{n-1}(x_j)) \\ &= \textstyle\sum_{j=1}^n w_j r_{n-1}(x_j)\end{aligned} \tag{5.6.17}$$

since $\psi_n(x_j) = 0$ for $j = 1, 2, \ldots, n$. We now combine (5.6.15), (5.6.16) and (5.6.17), to arrive at the completion of the proof of Lemma 5.6.3.

Theorem 5.6.4 *Let f be any function such that $(f,1) < \infty$, and such that $f^{(2n)}(x)$ exists on (a,b). Then*

$$\int_a^b w(x)f(x)dx = \sum_{j=1}^n w_j f(x_j) + \frac{f^{(2n)}(\zeta)}{(2n)!\,k_n^2} \tag{5.6.18}$$

where $\zeta \in (a,b)$, and k_n is defined in (5.5.4).

Proof: We recall (Chapter 4) the Lagrange-Hermite interpolation polynomial $P_{2n-1}(x)$ of degree $2n-1$ in x that satisfies

$$\begin{aligned}P_{2n-1}(x_j) &= f(x_j) \\ P_{2n-1}(x_j) &= f'(x_j)\end{aligned} \tag{5.6.19}$$

for $j = 1, 2, \ldots, n$, and also

$$f(x) - P_{2n-1}(x) = \frac{f^{(2n)}(\zeta)}{(2n)!\,k_n^2}\ \psi_n^2(x)\,, \tag{5.6.20}$$

where $\zeta = \zeta(x) \in (a, b)$ if $x \in [a, b]$. Upon setting

$$\epsilon(g) = \int_a^b w(x)g(x)dx - \sum_{j=1}^{n} w_j g(x_j) \tag{5.6.21}$$

and noting that $\epsilon(P_{2n-1}) = 0$, by Lemma 5.6.3, we obtain

$$\epsilon(f) = \epsilon(f) - \epsilon(P_{2n-1}) = \epsilon(f - P_{2n-1}) = \epsilon\left(\frac{f^{(2n)}(\zeta(x))}{(2n)\,!\,k_n^2}\right)\ \psi_n^2(x)\,. \tag{5.6.22}$$

Moreover, since $(\psi_n, \psi_n) = 1$ for $j = 1, 2, \ldots, n$, it follows that

$$\begin{aligned} \epsilon(f) &= \int_a^b w(x)\frac{f^{(2n)}(\zeta(x))}{(2n)\,!\,k_n^2}\ \psi_n^2(x)dx \\ &= \frac{f^{(2n)}(\zeta)}{(2n)\,!\,k_n^2}\int_a^b w(x)\psi_n^2(x)dx \qquad (5.6.23) \\ &= \tfrac{f^{(2n)}(\zeta)}{(2n)\,!\,k_n^2}\,,\ \zeta \in (a, b) \end{aligned}$$

using the result $(\psi_n, \psi_n) = 1$. This completes the proof of Theorem 5.6.4??.

When constructing quadrature formulas, it suffices by Sec. 5.3 to restrict ourselves to the intervals $(0, 1)$ or $(-1, 1)$ in the case when a and b are finite. For the case of the integral $\int_a^\infty f(y)dy$, we set $y = a + x$, and thus reduce to another standard interval $(0, \infty)$, while if $a = -\infty$ and $b = \infty$, we leave the interval unchanged.

Example 5.6.5 **(Gauss-Legendre Quadrature)** Here $w(x) = 1$, $a = -1$, $b = 1$. The orthogonal polynomials are the Legendre polynomials. The zeros $x_j^{(n)}$ and the corresponding weights $w_j^{(n)}$ are given in Table 5.6.1. The explicit expression for n-point Gauss-Legendre quadrature with error is

$$\int_{-1}^{1} f(x)dx = \sum_{j=1}^{n} w_j^{(n)} f(x_j^{(n)}) + \frac{2^{(2n+1)}(n\,!)^4 f^{(2n)}(\zeta)}{(2n+1)[(2n)\,!]^3} \tag{5.6.24}$$

where $\zeta \in (-1, 1)$.

Table 5.6.1: Abscissas and Weights for Gaussian-Legendre Integration

$x_j^{(n)}$		$w_j^{(n)}$
	$n=2$	
.577350269		1.000000000
	$n=3$	
.774596669		.555555556
.000000000		.888888889
.861136311		.347854845
.339981043		.652145155
	$n=5$	
.906179846		.236926895
.538469310		.478628670
.000000000		.568888899
	$n=6$	
.932469514		.171324452
.661209386		.360761573
.238619186		.467913934

Example 5.6.6 **(Gauss-Chebyshev Quadrature)** Here $a=-1$, $b=1$, and $w(x)=(1-x^2)^{-1/2}$. The orthogonal polynomials are the Chebyshev polynomials, $T_n(x)=\cos n\theta$, where $x=\cos\theta$. The zeros $x^{(n)}$ and the weights $w_j^{(n)}$ can be explicitly expressed:

$$\begin{aligned} x_j^{(n)} &= \cos\left[\tfrac{(2j-1)\pi}{2n}\right] \\ w_k^{(n)} &= \tfrac{\pi}{n}\,. \end{aligned} \tag{5.6.25}$$

Thus it may be shown that if $f\in C^{(2n)}[-1,1]$,

$$\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}\,dx=\frac{\pi}{n}\sum_{j=1}^{n}f\left(\cos\left[\frac{2j-1}{2n}\,\pi\right]\right)+\frac{\pi}{2^{2n-1}(2n)!}\,f^{(2n)}(\zeta) \tag{5.6.26}$$

where $\zeta\in(-1,1)$. This is the Gauss-Chebyshev quadrature based on the Chebyshev polynomials of the first kind $T_n(x)=\cos(n\theta)$, $\cos\theta=x$.

Example 5.6.7 **(Gauss-Laguerre Quadrature)** In this case $a=0$, $b=\infty$, and $w(x)=e^{-x}$. The orthogonal polynomials are the Laguerre polynomials. The zeros $x_j^{(n)}$ and the corresponding weights $w_j^{(n)}$ are given in Table 5.6.2 for several values of n. The explicit formula for Gauss-Laguerre quadrature with error is

$$\int_0^\infty e^{-x} f(x)dx = \sum_{j=1}^{n} w_j^{(n)} f(x_j^{(n)}) + \frac{(n\,!)^2}{(2n)\,!} f^{(2n)}(\zeta)\,, \tag{5.6.27}$$

where $\zeta \in (0, \infty)$.

Example 5.6.8 **(Gauss-Hermite Quadrature)** Here $a = -\infty$, $b = \infty$, and $w(x) = e^{-x^2}$. The orthogonal polynomials are the Hermite polynomials. The explicit form of Gauss-Hermite quadrature with error is

$$\int_{-\infty}^\infty e^{-x^2} f(x)dx = \sum_{j=1}^{n} w_j^{(n)} f(x_j^{(n)}) + \frac{\sqrt{\pi}\, n\,!}{2^n (2n)\,!} f^{(2n)}(\zeta) \tag{5.6.28}$$

where $\zeta \in (-\infty, \infty)$. The zeros $x_j^{(n)}$ and the corresponding weights $w_j^{(n)}$ are given in Table 5.6.3 for the first few values of n.

Tables exist for a number of different types of quadrature formulas. We refer the reader to Davis and Robinowitz [1], Strond and Secrest [2], and to the National Bureau of Standards Handbook of Mathematical Functions [3] for these. Recently, there has been considerable emphasis on the replacement of tables by efficient algorithms that generate the zeros and weights of Gaussian quadrature; here we refer the reader to Gautschi [4].

Example 6: Evaluate

$$I = \int_0^1 e^y dy\,.$$

If we use the trapezoidal formula, we obtain

$$I = \frac{1}{n}\left[\frac{1}{2} + e^{1/n} + e^{2/n} + \cdots + e^{(n-1)/n} + \frac{1}{2}\,e\right] + \epsilon$$

where $\epsilon = -e^{\zeta}/(12n^2)$, for some $\zeta \in (0,1)$. Thus the magnitude of the error is at least $1/(12n^2)$.

Transforming the integral to $(-1,1)$, we also have

$$I = \frac{1}{2}\,e^{1/2} \int_{-1}^1 e^{x/2} dx\,.$$

Using Gauss-Legendre quadrature to evaluate the integral, we get

$$I = \frac{1}{2}\,e^{1/2} \sum_{j=1}^{n} w_j^{(n)} e^{x_j^{(n)}/2} + \epsilon(e^{x/2})$$

Table 5.6.2: Abscissas and Weights for Gaussian–Laguerre Integration

$x_j^{(n)}$	$w_j^{(n)}$
$n=2$	
0.585786437	0.853553390
(1)0.341421356	0.146446609
$n=3$	
0.415774556	0.711093009
(1)0.229428036	0.278517733
(1)0.628994508	(-1)0.103892565
$n=4$	
0.322547689	0.603154104
(1)0.174576110	0.357418592
(1)0.453662029	(-1)0.388879085
(1)0.939507091	(-3)0.539294705
$n=5$	
0.263560319	0.521755610
(1)0.141340305	0.398666811
(1)0.359642577	(-1)0.759424496
(1)0.708581000	(-2)0.361175867
(2)0.126408008	(-4)0.233699723
$n=6$	
0.222846604	0.458964673
(1)0.118893210	0.417000830
(1)0.299273632	0.113373382
(1)0.577514356	(-1)0.103991974
(1)0.983746741	(-3)0.261017202
(2)0.159828739	(-6)0.898547906

$(d)0.a_1a_2 = .a_1a_2 \times 10^d$

Table 5.6.3: Abscissas and Weights for Gaussian–Hermite Integration

$x_j^{(n)}$		$w_j^{(n)}$
	$n=2$	
0.707106781		0.886226925
	$n=3$	
(1)0.122474487		0.295408975
0.000000000		(1)0.118163590
	$n=4$	
(1)0.165068012		(-1)0.813128354
0.524647623		0.804914090
	$n=5$	
(1)0.202018287		(-1)0.199532420
0.958572464		0.393619323
0.000000000		0.945308720
	$n=6$	
(1)0.235060497		(-2)0.453000990
(1)0.133584907		0.157067320
0.436077411		0.724629595
$(d)0.a_1a_2 = .a_1a_2 \times 10^d$		

where, using (**??**)

$$\epsilon(e^{x/2}) = \frac{1}{2}\, e^{1/2} \cdot \frac{2^{2n+1}(n\,!\,)^4}{(2n+1)[(2n)\,!\,]^3} \left(\frac{d^{2n}}{dx^{2n}} (e^{x/2}) \right) \bigg|_{x=\zeta},$$

and where $\zeta \in (-1,1)$. Since $(e^{x/2})^{(2n)} = (1/2)^{2n} e^{x/2}$ we find that the error of Gaussian quadrature is

$$\begin{aligned} \epsilon(e^{x/2}) &= \frac{e^{(\xi+1)/2}(n\,!\,)^4}{(2n+1)[(2n)\,!\,]^3} \\ &\leq e \left(\frac{2\pi}{n}\right)^{1/2} \left(\frac{e}{8n}\right)^{2n} \end{aligned}$$

where Sterling's formula is used to obtain the inequality in (5.6.24). A comparison of the error bounds are listed in Table 5.6.4.

The above example illustrates the rapid convergence of Gaussian quadrature when applied to a suitable integrand. Rapid convergence occurs for

Table 5.6.4: Error Bounds for Example ??

n	Lower Bound for Trapezoidal Error	Upper Bound for Gauss-Legendre Error
2	2.1×10^{-2}	4.0×10^{-3}
5	3.3×10^{-3}	6.4×10^{-12}
10	8.3×10^{-4}	9.1×10^{-30}

those functions whose $2n$'th derivative exists for every positive integer n, and grows slowly as a function of n. The function $e^{x/2}$ is of this type.

An integral

$$I = \int_a^b F(y)dy \tag{5.6.29}$$

can be effectively approximated by Gaussian quadrature if the following conditions are met.

(i) It is possible to write

$$I = \int_a^b w(x)f(x)dx \tag{5.6.30}$$

where $w(x) > 0$ on (a, b), and where the maximum value of the $2n$'th derivative of f on $[a, b]$ grows slowly as a function of n. This condition is met, for example, if the series

$$f(x) = \sum_{n=0}^{\infty} c_n \left(x - \frac{a+b}{2}\right)^n \tag{5.6.31}$$

converges for $\left|x - \frac{a+b}{2}\right| > (b-a)/2$.

(ii) The zeros $x_j^{(n)}$ and the corresponding weights $w_j^{(n)}$ are already tabulated for different values of n, or else they can be easily computed.

Example 7: An Oscillatory Integrand:

$$\int_0^{2\pi} x \sin(\ell x) \cos(x) dx \, .$$

Table 5.6.5: Results for Example 7??

ℓ	Exact	32-Point Legendre Quadrature
1	-1.57079	-1.57079
2	-4.18879	-4.18428
4	-1.67552	-1.67565
10	-.63467	-.63402
20	-.31495	-1.20925
30	-.20967	-1.58222

This integral was approximated by 32-point Gauss-Legendre quadrature. The results are given in Table 5.6.5.

Problems

5.6.1 Show that the sequence of polynomials $\{\psi_n\}_{n=0}^{\infty}$ constructed in (5.5.5) satisfies (5.5.3). Hint: Note first of all that $\{\psi_n\}_{n=0}^{N}$ is an independent set. For our inductive hypothesis assume that $\{\psi_n\}_{n=0}^{N-1}$ are orthonormal (they clearly satisfy $(\psi_n, \psi_n) = 1$ in (5.5.3)). Now compute $(\psi_N, \psi_j) = (\theta_N, \theta_N)^{-1/2}(\theta_N, \psi_j)$ for $j < N$ using (5.5.6) and independent of $\{\psi_n\}_{n=0}^{N}$.

5.6.2 Show that if $\pi_j(x)$ is an arbitrary polynomial of degree $j < n$ then $(\psi_n, \pi_j) = 0$.

5.6.3 Show that the sequence $\{\theta_n\}_{n=0}^{\infty}$ satisfies the recurrence relation

$$\begin{aligned} \theta_0 &\equiv 1 \qquad\qquad \theta_1(x) = x - \beta_1 \\ \theta_n(x) &= (x - \beta_n)\theta_{n-1}(x) - \gamma_{n-1}\theta_{n-2}(x) \ , \quad n \geq 2 \end{aligned}$$

where

$$\beta_n = \frac{(x\,\theta_{n-1}, \theta_{n-1})}{(\theta_{n-1}, \theta_{n-1})}\ , \ \gamma_{n-1} = \frac{(x\,\theta_{n-1}, \theta_{n-2})}{(\theta_{n-2}, \theta_{n-2})}\ .$$

Hint: Using the notation in (5.6.3) write

$$\begin{aligned} \theta_n(x) - x\,\theta_{n-1}(x) &= -\beta_n\theta_{n-1}(x) - \gamma_{n-2}\theta_{n-2} + c_{n-3}\theta_{n-3} \\ &\quad + \cdots + c_1\theta_1(x) + c_0\theta_0(x) \end{aligned}$$

and use the orthogonality of $\{\theta_j\}_{j=0}^{n-3}$ to show that $c_j = 0$ for $j = 0, l \ldots, n-3$. Now use the orthogonality of θ_{n-1} and θ_n to find β_n and γ_{n-1}.

5.6.4 If $[a, b] = [-1, 1]$ and $w(x) = 1$ in (5.6.3) the resulting orthonormal polynomials are called the Legendre polynomials. Use Problem 5.6.3 to find the first five Legendre polynomials. Verify that the roots $x_j^{(n)}$ in Table 5.6.1 are correct $(n = 2, 3, 4, 5)$.

5.6.5 Verify Lemma 5.6.2.

5.6.6 How many evaluation points are required to evaluate $\int_{-1}^{1}(1-x^2)^{-1/2}e^x dx$ accurate to eight decimals using the Gauss-Chebyshev quadrature (5.6.26)?

5.6.7 Prove that the weights $w_j^{(n)}$ in (5.6.18) are given by

$$w_j^{(n)} = -\frac{k_{n+1}}{k_n \psi_{n+1}(x_j)\psi_n'(x_j)}$$

where k_n is defined in (5.6.4).

5.6.8 If $w(x) = (1 - x^2)^{1/2}$, $a = -1$, $b = +1$ and $f \in C^{2n}[-1, 1]$ then

$$\int_{-1}^{1}(1 - x^2)^{1/2} f(x)dx = \sum_{j=1}^{n} w_j^{(n)} f(x_j^{(n)}) + \frac{\pi}{2^{2n+1}(2n)!}\, f^{2n}(\xi)$$

for some $\zeta \in (-1, 1)$, where

$$x_j^{(n)} = \cos\left(\frac{j\pi}{n+1}\right) \,,\; w_j^{(n)} = \pi \sin^2\left(\frac{j\pi}{n+1}\right) .$$

This is the Gauss-Chebyshev quadrature based on the Chebyshev polynomials of the second kind $u_n(x) = \sin((n + 1)\theta)/\sin\theta$, $\cos\theta = x$.

5.6.9 Show that the quadrature

$$\int_{1}^{3} f(x)dx = \frac{5}{9} f(2 - \sqrt{315}) + \frac{8}{9} f(2) + \frac{5}{9} f(2 + \sqrt{315})$$

is exact when f is a polynomial of degree 5.

5.6.10 Let $\{\psi_n\}$ be defined as in (5.6.5). Prove that

$$\sum_{j=1}^{n} \int_{a}^{b} w(x) \left[\frac{\psi_n(x)}{(x - x_j^{(n)})\psi_n'(x_j^{(n)})}\right]^2 dx = \sum_{j=1}^{n} w_j^{(n)} = \int_{a}^{b} w(x)dx \,.$$

5.6.11 Find x_j and w_j so that the formula

$$\int_0^1 x^{2/3} f(x)dx = \sum_{j=1}^{5} w_j f(x_j)$$

is exact when f is a polynomial of degree nine.

5.7 Quadratures Based on the Trapezoidal Rule

Let us briefly recall various features of the quadrature formulas constructed in the previous sections. The trapezoidal and mid-ordinate formulas are very simple to use on a computer. Using (5.3.11), it was possible to use previously computed evaluation points by halving the step-size h. A disadvantage of the trapezoidal and mid-ordinate rules is that they converge so slowly, even for smooth functions, that it is usually impossible to achieve more than five significant figures of accuracy using eight significant figure arithmetic. This situation is often remedied by using the Romberg algorithm; however, when singularities are present at endpoints of integration, the Romberg method, as well as the trapezoidal and mid-ordinate methods, usually fails.

Due to the relatively few points they require, the Gaussian formulas often yield very accurate results in the case when the integrand is smooth. Their disadvantage is that to increase the accuracy, we cannot use previously computed evaluation points, and it is thus inconvenient to obtain an estimate of the error on a posteriori basis. Furthermore, in the case when singularities are present at an endpoint of integration, i.e., when in the approximation

$$\int_a^b w(x)f(x)dx \simeq \sum_{j=1}^{n} w_j^{(n)} f(x_j^{(n)})$$

the function f also has singularities the Gaussian formulas are not much better than the mid-ordinate formulas.

Ideally, one would like formulas that are simple to apply on a computer, for which one can use previously computed evaluation points via a device such as that of (5.3.11), which converge as rapidly as the trapezoidal formula in Table 5.4.2, so that we can easily estimate the error a posteriori, and which are usually insensitive to singularities located at endpoints of integration. While there are no universal-type quadrature formulas that yield accurate results for all integrands (c.f. [5] and Problems 5.7.9–5.7.14), the formulas that we construct in this section do have the desirable features that we have just described. We only list three formulas; further generalizations and the effectiveness of the formulas when singularities are present at endpoints of integration can be found in the literature [6].

Theorem 5.7.1 *(a) Let $f^{(k)}(x)$ exist on the interval $(-\infty, \infty)$, and let*

$$\int_{-\infty}^{\infty} |f^{(k)}(x)| dx < \infty \tag{5.7.1}$$

for $k = 0, 1, 2, 3, \ldots$. Then for every $\alpha > 0$

$$\int_{-\infty}^{\infty} f(x) dx - h \sum_{k=-\infty}^{\infty} f(kh) = 0(h^{\alpha}) \tag{5.7.2}$$

as $h \to 0$.

(b) Let $F^{(k)}(x)$ exist on the interval $(0, \infty)$, and let

$$\int_{0}^{\infty} y^k |F^{(k)}(y)| dy < \infty \tag{5.7.3}$$

for $k = 0, 1, 2, 3, \ldots$. Then for every $\alpha > 0$

$$\int_{0}^{\infty} F(y) dy - h \sum_{k=-\infty}^{\infty} e^{kh} F(e^{kh}) = 0(h^{\alpha}) \tag{5.7.4}$$

as $h \to 0$.

(c) Let $\mathcal{F}^{(k)}(x)$ exist on the interval $(-1, 1)$, and let

$$\int_{-1}^{1} (1 - x^2)^k |\mathcal{F}^{(k)}(x)| dx < \infty \tag{5.7.5}$$

for $k = 0, 1, 2, 3, \ldots$. Then for every $\alpha > 0$

$$\int_{-1}^{1} \mathcal{F}(x) dx - h \sum_{k=-\infty}^{\infty} \frac{2e^{kh}}{(1 + e^{kh})^2} \, \mathcal{F}\left(\frac{e^{kh} - 1}{e^{kh} + 1}\right) = 0(h^{\alpha}) \tag{5.7.6}$$

as $h \to 0$.

Before proving this theorem, we state some properties of the formulas, and we describe an effective algorithm for using them. Each of the formulas takes the form

$$\int_a^b g(x)dx \cong h \sum_{k=-\infty}^{\infty} w_k(h) g(x_k(h)), \tag{5.7.7}$$

where, for example, in formula (5.7.6), $a = -1$, $b = 1$, $w_k(h) = 2e^{kh}/(1 + e^{kh})^2$, and $x_k(h) = (e^{kh} - 1)/(e^{kh} + 1)$. For each of the formulas $w_{2k}(h/2) = w_k(h)$, $x_{2k}(h/2) = x_k(h)$. These remarks lead us to the following algorithm for evaluating $\int_a^b g(x)dx$ to within ϵ, where $\epsilon > 0$ is given.

Algorithm 5.7.2 Quadrature Based on the Trapezoidal Rule

Pick c and d, $a < c < d < b$, such that

$$\int_a^c |g(x)|dx < \frac{\epsilon}{3} \quad \text{and} \quad \int_d^b |g(x)|dx < \frac{\epsilon}{3}$$

Pick $h > 0$.

Evaluate $\quad T_h(g) \equiv h \sum_{x_k(h) \in (c,d)} w_k(h) g(x_k(h))$

Evaluate $\quad M_h(g) \equiv h \sum_{x_{2k-1}(h/2) \in (c,d)} w_{2k-1}(h/2) g(x_{2k-1}(h/2))$

Note $\quad T_{h/2}(h) = \frac{1}{2}\left[T_h(g) + M_h(g)\right]$

If $|T_{h/2}(g) - T_h(g)| < \epsilon/3$

then accept $T_{h/2}(g)$

else

$h \leftarrow h/2$

Return

Proof of Theorem 5.7.1

(a) This is a direct consequence of Corollary 5.4.7

(b) Here we set

$$x = e^t \tag{5.7.8}$$

in the integral in (5.6.4), and then apply the result (5.7.2), to get

$$\int_0^\infty F(x)dx = \int_{-\infty}^\infty e^t F(e^t) = h \sum_{k=-\infty}^\infty e^{kh} F(e^{kh}) = 0(h^\alpha) \tag{5.7.9}$$

as $h \to 0$, for all $\alpha > 0$, provided that

$$\int_{-\infty}^\infty |(e^t F(e^t))^{(k)}| dt < \infty \tag{5.7.10}$$

for $k = 0, 1, 2, 3, \ldots$. Now it is readily verified by induction (see Problem 5.7.4) that there are constants $a_{k\ell}$, $\ell = 0, 1, \ldots, k-1$, such that

$$(e^t F(e^t))^{(k)} = (e^{(k+1)t}) F^{(k)}(e^t) + \sum_{\ell=0}^{k-1} a_{k\ell} (e^{(\ell+1)t}) F^{(\ell)}(e^t) \tag{5.7.11}$$

where the sum on the right is zero if $k = 0$. Hence, if

$$\int_0^\infty y^k |F^{(k)}(y)| dy = \int_{-\infty}^\infty e^{(k+1)t} |F^{(k)}(e^t)| dt < \infty \tag{5.7.12}$$

for $k = 0, 1, 2, 3, \ldots$, then from (5.7.11) it follows that (5.7.10) is satisfied.

(c) For this case we set

$$x = \tanh \frac{1}{2} w \tag{5.7.13}$$

in the integral on the right of (5.7.6), and substitute into (5.6.2) to get

$$\begin{aligned} \int_{-1}^1 \mathcal{F}(x) dx &= \int_{-\infty}^\infty \frac{1}{2} \operatorname{sech}^2 \left(\frac{1}{2} w \right) \mathcal{F} \left(\tanh \frac{1}{2} w \right) dw \\ &= h \sum_{k=-\infty}^\infty \left(\frac{2e^{kh}}{(1+e^{kh})^2} \right) \mathcal{F} \left(\frac{e^{kh}-1}{e^{kh}+1} \right) + 0(h^\alpha) \end{aligned} \tag{5.7.14}$$

as $h \to 0$, for all $\alpha > 0$, if the function $f(w) = (1/2)\operatorname{sech}^2(w/2)\mathcal{F}(\tanh(w/2))$ satisfies (5.7.1) for $k = 0, 1, 2, \ldots$. Now we can verify by induction (see Problems 5.7.5, 5.7.6 and 5.7.7) that there are functions $b_{k\ell}(x)$,

$\ell = 0, 1, \ldots, k-1$, which are uniformly bounded for $-\infty \le x \le \infty$, such that

$$((1/2)\text{sech}^2(x/2)\psi(\tanh x/2)) = \frac{\text{sech}^2 x/2}{2}\,\psi^{(k)}(\tanh(x/2)) + \sum_{\ell=0}^{k-1} b_{k\ell}(x)\text{sech}^{2\ell+2}(x/2)\psi^{(\ell)}(\tanh x/2) \quad . \tag{5.7.15}$$

Hence if

$$\int_{-1}^{1}(1 - x^2)^{(k)}\mathcal{F}^{(k)}(x)dx = \frac{1}{2}\int_{-\infty}^{\infty}|(\text{sech}^2(x/2))^{(k+1)}|\,|\mathcal{F}^{(k)}(\tanh(x/2))|dx < \infty \tag{5.7.16}$$

for $k = 0, 1, 2, \ldots$, then (5.7.5) is satisfied.

Example 5.7.3 We approximate

$$\int_4^{\infty}\frac{e^{-x^2}}{\sqrt{2\pi}}\,dx = .54508\ 23808\ 67\times 10^{-8}$$

by use of (5.7.4). Here we first replace x by $4+y$, and then apply (5.2.4) in the Algorithm 5.7.2. The results are contained in Table 5.7.

Table 5.7.1:

h	Number of Points	Results $\times 10^8$
$\log 2$	23	.545080
$\frac{1}{2}\log 2$	79	.54508238081
$\frac{1}{4}\log 2$	149	.545082380867

Example 2: We have already seen the effectiveness of Gaussian quadrature (see Example 6 in Section 6) for evaluating an integral of the form

$$I = \int_{-1}^{1} e^x\,dx\ .$$

Table 5.7.2: Comparison of Number of Points

Trap Rule	Gauss	6.6	Accuracy
900	4	21	6.55 – 7
912,870	7	83	4.29 – 13
9×10^{11}	9	329	1.84 – 25

We evaluate this integral using the trapezoidal rule, the Gauss-Legendre rule, and the formula of (5.7.6), and we compare the number of points required to achieve the same accuracy. The results are contained in Table 5.7.

Thus the formula of (5.7.6) is not as accurate as Gaussian quadrature for some integrands, although in the case of this particular example, it is considerably more accurate than the trapezoidal rule.

We do not want to leave the reader with the impression that the formulas derived in this section are universal, since formulas that converge rapidly for all integrands do not exist (see Problems 5.7.9–5.7.14). For example, when evaluating

$$\int_{-\infty}^{\infty} (1+x^2)^{-1} dx$$

by (5.7.2) the error can be shown to be $4\pi e^{-2\pi/h}/(1-e^{-2\pi/h})$. Although the error converges to zero very rapidly as a function of h, many evaluation points are required, since the integrand converges to zero slowly as $x \to \pm\infty$. Note, however, that (5.5.4) works well for evaluating $\int_0^\infty (1+x^{-2})dx$. Although $\int_a^\infty [(\sin x)/x]dx$ exists for every finite a, we cannot evaluate this integral by use of the formula (5.7.4), since the resulting infinite sum is divergent for every $h > 0$. Finally, the formula of (5.7.6) fails to converge rapidly for the valuation of $\int_0^1 x \sin(1/x) dx$.

Problems

5.7.1 Let $\alpha \in (0, 2)$. Find constants $c_1 > 0$ and $c_2 > 0$, such that

$$\frac{c_1}{n^\alpha} < \int_0^1 x^{\alpha-1} dx - \frac{1}{n} \sum_{k=1}^{n} \left(\frac{2k-1}{2n} \right)^{\alpha-1} < \frac{2}{n^\alpha}\,, \quad n = 1, 2, 3, \ldots .$$

5.7.2 Find $\int_{-1}^1 (1-x^2)^{-1/2} dx$ accurate to six significant figures using the formula (5.7.6) in the Algorithm 7.2.

5.7.3 Evaluate $\int_0^\infty y^{-1/2}(1+y)^{-1}dy$ correct to six significant figures, using (5.7.4) in Algorithm 5.2.

5.7.4 Verify (5.7.11) by showing that the constants $a_{k\ell}$ satisfy the recurrence relations

$$\begin{aligned} a_{k+1,\ell} &= (\ell+1)a_{k\ell} + a_{k,\ell-1} \ , \quad \ell = 0,1,\ldots,k-1 \\ a_{k+1,k} &= k+1+a_{k,k-1} \ . \end{aligned}$$

5.7.5 Verify that there are constants α_{jk} such that

$$(\tanh(x/2))^{(2k-1)} = \sum_{j=1}^{2k-1} \alpha_{jk}\mathrm{sech}^{2j}(x/2) \ , \quad k = 1,2,3,\ldots .$$

5.7.6 Verify that these are constants β_{jk} such that

$$(\tanh(x/2))^{(2k)} = \tanh(x/2)\sum_{j=0}^{2k} \beta_{jk}\mathrm{sech}^{2j}(x/2) \ , \quad k = 0,1,2,\ldots .$$

5.7.7 Use the results of Problems 5.7.5 and 5.7.6 to verify (5.7.15), where $b_{k\ell}^{(j)}(x)$ is uniformly bounded for $-\infty \le x \le \infty$.

5.7.8 Show that if $\int_{-\infty}^{\infty} |f^{(k)}(x)|dx < \infty$ for $h = 0,1,2,\ldots$, then $f^{(k)}(x) \to 0$ as $x \to \pm\infty$.

5.7.9 Let $Q_n(f)$ be defined by

$$Q_n(f) = \frac{1}{2^{n-1}}\left[f\left(\frac{1}{2^n}\right) + f\left(\frac{3}{2^n}\right) + \cdots + f\left(\frac{2^n-1}{2^n}\right)\right] .$$

Prove that $Q_n(f) \to \int_0^1 f(x)dx$ as $n \to \infty$ for all $f \in C[0,1]$.

5.7.10 Let π_n denote the set of points $\left\{\frac{2k-1}{2^n}\right\}_{k=1}^{2^{n-1}}$, $n = 1,2,\ldots$ and let $s_n(x)$ denote the linear spline that is 1 for $x \in \pi_n$, and zero for all points halfway between two consecutive points of π_n as well as for $x = 0$ and $x = 1$. Prove that if $\sum_{n=1}^{\infty} |\alpha_n| < \infty$, then the function f defined by $f(x) = \sum_{n=1}^{\infty} \alpha_n s_n(x)$ is in $C[0,1]$.

5.7.11 In the notation of Problems 5.7.9 and 5.7.10, prove that $\int_0^1 s_n(x)dx = 1/2$, $n = 1, 2, 3, \ldots$, and also that $Q_i(s_j) = 1/2$ if $j < i$, 1 if $j = i$, and 0 if $j > i$.

5.7.12 Let $a_n > 0$ for $n = 1, 2, \ldots$ and $a_1 > a_2 > \cdots$ with $a_n \to 0$ as $n \to \infty$. Let $\alpha_1 = 0$ and α_n $(n \geq 2)$ be defined by $\alpha_{n+1} = (1/2)\alpha_n + a_n - a_{n+1}$, $n = 1, 2, \ldots$. Prove that if f is defined as in ???6.10 then the function $f = \sum_{n=1}^{\infty} \alpha_n s_n(x) \in C[0, 1]$.

5.7.13 Prove that if f is defined as in Problem **??**, and Q_n as in Problem 5.6.9, then

$$\int_0^1 f(x)dx - Q_n(f) = a_n \ , \quad n = 1, 2, 3 \ldots .$$

5.7.14 Prove that the function f defined in Problem **??** satisfies $\|f(x)\| \leq 2a_1$, $0 \leq x \leq 1$.

Chapter 6

Asymptotic Methods

Introduction

Often an integral or solution of a differential equation is expressed in the form $f(x,n)$ or $g(x)$ where a closed (or simple) form for f or g is not obtainable. A classic example is the integral that represents probability given a normal distribution. This example is part of a select group used so often that tabular values are listed in many handbooks. If the problem under study does not fall into this category and information regarding the limiting behavior of a function is sought then we are in that branch of mathematics called asymptotic analysis. Briefly, for the functions f and g we see simpler functions $f_0(x,n)$ or $g_0(x)$ satisfying

$$f(x,n) = f_0(x,n)[1+o(1)]\ , \quad x \text{ or } n \to \infty$$

or

$$g(x) = g_0(x)[1+o(1)]\ , \quad x \to \infty\,.$$

The functions f_0 and g_0 are called asymptotic estimates of f and g, respectively.

The goal of this chapter is twofold. First, we are concerned with ascertaining f_0 and g_0 and, second, with bounding the error $\epsilon(x)$ where for example

$$g(x) = g_0(x)[1+\epsilon(x)]\,.$$

One method for bounding ϵ is to express ϵ in reciprocal powers of x. The result

$$\epsilon(x) = \frac{\epsilon_1}{x} + \frac{\epsilon_2}{x^2} + \cdots + \frac{\epsilon_{k-1}}{x^{k-1}} + \cdots$$

is called an asymptotic series. This series may not converge, and upon truncating the series after a finite number of terms we may again wish to bound the error, e.g. to bound $\eta_k(x)$ in

$$g(x) = g_0(x)\left[1 + \frac{\epsilon_1}{x} + \cdots + \frac{\epsilon_{k-1}}{x^{k-1}} + \eta_k(x)\right].$$

This chapter differs from most accessible texts in two ways. First, the chapter is introductory and elementary in that only real variable methods are used. Second, attention is restricted to those parts of asymptotic analysis that have a direct relation to the subject matter of the text. In particular, the methods presented here complement and extend the analysis in Chapters 5 and 7.

6.1 Integration by Parts

To fix ideas define the exponential integral

$$\mathrm{Ei}(x) = \int_x^\infty \frac{e^{-t}}{t}\,dt \tag{6.1.1}$$

where x is a large positive parameter. If one integration by parts is performed in the integral in (6.1.1) then

$$\mathrm{Ei}(x) = \frac{e^{-x}}{x} - \int_x^\infty \frac{e^{-t}}{t^2}\,dt \tag{6.1.2}$$

and the first term on the right-hand side of (6.1.2) is called an asymptotic estimate of the integral. Continuing integration by parts in (6.1.2) leads to the expression

$$\begin{aligned}\mathrm{Ei}(x) &= e^{-x}\left[\frac{1}{x} - \frac{1!}{x^2} + \frac{2!}{x^3} - \cdots + \frac{(-1)^{n-1}(n-1)!}{x^n}\right] + R_n(x) \\ &\equiv e^{-x}\left[\sum_{k=0}^{n-1}\frac{(-1)^k k!}{x^{k+1}}\right] + R_n(x)\end{aligned} \tag{6.1.3}$$

where

$$R_n(x) = (-1)^n n!\int_x^\infty \frac{e^{-t}}{t^{n+1}}\,dt\,. \tag{6.1.4}$$

Table 6.1.1: Ei(5) = .00114829

n	$\mathrm{Ei}^n(5)$	$\mathrm{Ei}(5) - \mathrm{Ei}^{(n)}(5)$
1	.00134758	-.000199
2	.00107807	+.000070
3	.00118588	-.000037
4	.00112119	+.000027
5	.00117294	-.000024
6	.00112119	+.000027
7	.00118329	-.000035
8	.00109635	+.000052
9	.00123545	-.000087
10	.00098508	+.000163
11	.00148583	-.000338

$\mathrm{Ei}^{(n)}(5) \equiv e^{-x}\left[\frac{1}{x} - \frac{1\,!}{x^2} + \frac{2\,!}{x^3} - \cdots + \frac{(n-1)\,!}{x^n}\right]$

Consider evaluating Ei(5) via the formula in (6.1.3). These results are displayed in Table 6.1.1.

A graph of the error (using the right-hand column in Table 6.1.1) has the oscillatory
appearance displayed in Figure 6.1.1.

A survey of the figure or the last column in Table 6.1.1 shows that as n increases up to 5 the error decreases, and as n increases past 5 there is a deterioration in the approximation. Indeed, the series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}(n-1)\,!}{x^n} \tag{6.1.5}$$

diverges for every finite x but at $n = 5$ the error in approximating Ei(5) by five terms of (6.1.5) gives the value of the function to four digits (compare to quadratures). The series in (6.1.5) is called an asymptotic series (in the powers $1/x$) and such series often provide a good (if limited) approximation to the expression generating them (in this case the function $\mathrm{Ei}(x)$).

To see why $n = 5$ provided the best accuracy consider

$$(-1)^n R_n(x) \leq \frac{n\,!}{x^{n+1}} \int_x^{\infty} e^{-t} dt = \frac{n\,! \, c^{-x}}{x^{n+1}} \tag{6.1.6}$$

Figure 6.1.1:

valid, using (6.1.4), for all $0 < x \leq t < \infty$. Hence the error in truncating the asymptotic series is bounded in magnitude by the first deleted term and has the sign of the first deleted term (hence the oscillatory graph in Figure ??6.1.2).

The ratio of the magnitude of the $(n+1)^{\text{st}}$ term to the nth term is

$$\frac{n\,!\,/x^{n+1}}{(n-1)\,!\,/x^{n}} = \frac{n}{x}$$

which for fixed x is less than one until $n = [|x|] \equiv$ integer part of x and thereafter increases without bound ($n = x = 5$ in Table 6.1.1).

This example illustrates the typical behavior of asymptotic series: Usually only a limited accuracy is possible for each x. While it is possible, in this example, to pick the optimal value of n for given x, in practice the choice of the optimal n is more difficult if the error term is complicated.

The preceeding example is a special case of integrals in the form

$$g(x) = \int_x^\infty e^{\alpha t^m} f(t) dt \,, \tag{6.1.7}$$

with $\alpha = -1$, $m = 1$ and $f(t) = 1/t$. Two other important special cases are the Probability distribution

$$Erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt \tag{6.1.8}$$

and the Fresnel integrals

$$C(x) + iS(x) = \int_x^\infty e^{it^2} dt \,. \tag{6.1.9}$$

One integration by parts in (6.1.8) gives

$$Erfc(x) = \frac{e^{-x^2}}{\sqrt{\pi}\, x} + \frac{1}{\pi} \int_x^\infty \frac{e^{-t^2}}{t^2} dt \,, \tag{6.1.10}$$

but, instead of continued integration by parts, consider the more general approach applied to (6.1.7). Define the change of variables

$$t = u^{1/m} \,, \quad u = v + x^m \tag{6.1.11}$$

so that (6.1.7) takes the form

$$\begin{aligned} g(x) &= \frac{1}{m} \int_{x^m}^\infty \frac{e^{\alpha u} f(u^{1/m})}{u^{(1-1/m)}} du \\ &= \frac{e^{\alpha x^m}}{m} \int_0^\infty \frac{e^{\alpha v} f((v + x^m)^{1/m})}{(v + x^m)^{(1-1/m)}} dv \,. \end{aligned} \tag{6.1.12}$$

Using (6.1.12) for (6.1.8) ($m = 2$, $\alpha = -1$ and $f(t) \equiv 1$) leads to the expression

$$\begin{aligned} Erfc(x) &= \frac{2e^{-x^2}}{\sqrt{\pi}\, 2} \int_0^\infty \frac{e^{-v}}{(v + x^2)^{1/2}} dv \\ &= \frac{e^{-x^2}}{\sqrt{\pi}\, x} \int_0^\infty \left(1 + \frac{v}{x^2}\right)^{-1/2} e^{-v} dv \\ &= \frac{e^{-x^2}}{\sqrt{\pi}\, x} \int_a^\infty \left\{ 1 + \sum_{k=1}^\infty \frac{(-1)^k (2k-1)!!}{(2k)!!\, x^{2k}} v^k \right\} e^{-v} dv \end{aligned} \tag{6.1.13}$$

$$
\begin{aligned}
&\doteq \frac{e^{-x^2}}{\sqrt{\pi}\,x}\left\{1+\sum_{k=1}^{\infty}\frac{(-1)^k(2k-1)\,!\,!}{(2k)\,!\,!\,x^{2k}}\,k\,!\right\} \\
&= \frac{e^{-x^2}}{\sqrt{\pi}\,x}\left\{1+\sum_{k=1}^{\infty}\frac{(-1)^k(2k-1)\,!\,!}{(2x^2)^k}\right\}
\end{aligned}
$$

where $(2k-1)\,!\,! = (2k-1)(2k-3)\cdots 1$, and the integral

$$\int_0^{\infty} v^k e^{-v}dv = k\,!$$

have been used in line following the $\doteq$. Notice that the first term on the right-hand side of (6.1.13) is the same as that in (6.1.10). Indeed the series in (6.1.13) is the same as that obtained from continued integration by parts in (6.1.10) (Problem????6.1). As in the series (6.1.5) the extreme right of (6.1.13) diverges for every positive x. This, of course, is in violent contrast to the integral defining $Erfc(x)$ which converges for every x. It is not hard to find the "breakdown" in (6.1.13). The series expansion of $(1+v/x^2)^{-1/2}$ converges only in $|v/x^2| < 1$ (Problem ???6.2) but the series development was used over the entire range $v \in [0,\infty)$. It is however the case that the extreme right of (6.1.13) is an asymptotic expansion of $Erfc(x)$. The following powerful lemma justifies the expansion in (6.1.13).

Lemma 6.1.1 "Watson's" *For $|t| \le R$, let*

$$f(t) = \sum_0^{\infty} a_n t^n \tag{6.1.14}$$

and for $|t| > R$, let

$$|f(t)| \le C\,e^{bt}\,. \tag{6.1.15}$$

If $m > -1$, then

$$
\begin{aligned}
\Big|\int_0^{\infty} e^{-xt}t^m f(t)dt \quad &- \quad \sum_{k=0}^{n-1}\frac{a_k\Gamma(k+m+1)}{x^{k+m+1}}\Big| \\
&\le \quad M\,\frac{\Gamma(m+n+1)}{x^{m+n+1}} + C\,e^{-xR}
\end{aligned}
\tag{6.1.16}
$$

where C and M are constants independent of m, x and R.

The proof of this theorem can be found in a number of texts and is omitted here since we carry out a similar argument in the proof of the method of Laplace in 6.3.

Truncation of the series obtained from substitution of (6.1.14) into the integral in (6.1.16) yields

$$\int_0^\infty e^{-xt} t^m \sum_{k=0}^{n-1} q_k t^k dt\,,$$

a partial sum, $f(t)$ being replaced by the first n lines of its Taylor series. The factor t^m in the integrand in (6.1.16) serves to cope with singularities at $t = 0$ that may arise as a result of a transformation of the type (6.1.11).

The first term on the right of (6.1.16) is proportional to the first term neglected in replacing the series

$$\sum_0^\infty a_k \frac{\Gamma(m+k+1)}{x^{m+k+1}} \quad \text{by} \quad \sum_0^{n-1} a_k \frac{\Gamma(m+k+1)}{x^{m+k+1}}\,.$$

Watson's lemma is not a convenient tool for estimating the error of the asymptotic approximation; its chief use is that it justifies the familiar procedure of approximating

$$\int_0^\infty e^{-xt} t^m f(t) dt$$

by expanding f in powers of t and performing termwise integration. The examples 1 to 6 below illustrate the use and limitations of Watson's lemma.

Example 6.1.2 Put $t = v/x^2$ in the second line of (6.1.13) then the last line of (6.1.13) is the result of Watson's lemma applied to $f(t) = (1+t)^{-1/2}$.

Example 6.1.3 For the integral in (6.1.19), using (6.1.12) it follows that

$$\begin{aligned} C(x) + iS(x) &= \int_x^\infty e^{it^2} dt \\ &= \frac{1}{2}\int_{x^2}^\infty \frac{e^{iu}}{u^{1/2}} du \\ &= \frac{e^{ix^2}}{2}\int_0^\infty \frac{e^{iv}}{(v+x^2)^{1/2}}\, dv\,. \end{aligned}$$

Watson's lemma is not directly applicable, since the integrals

$$\int_0^\infty e^{iv} v^k dv$$

do not exist.

Example 6.1.4 The ratio of the magnitude of successive terms in the expression (6.1.16) using Watson's lemma is

$$r_k = \left| \frac{a_{k+1}}{a_k} \frac{k+m+1}{x} \right| .$$

If $a_{k+1}/a_k \to \rho$ as $k \to \infty$, where ρ is finite, $r_k \to \infty$ as $k \to \infty$ for each finite x. That is, the asymptotic series

$$\sum_0^\infty a_k \frac{\Gamma(m+k+1)}{x^{m+k+1}} \tag{6.1.17}$$

diverges for each finite x. The series (6.1.17) converges for all sufficiently large x if $\lim r_k < \infty$, $k \to \infty$. In this latter case the series expansion (6.1.15) converges to $f(t)$ for all finite values of t.

Example 6.1.5 The integral

$$I(y) = \int_0^\infty e^{yt - 1/3\, t^3} dt \;, \quad y \to 0$$

can be replaced by an asymptotic series in $x = 1/y^3$ by means of Watson's lemma. Setting

$$yt = (3v)^{1/3} \;, \quad y = x^{-1/3}$$

we get

$$\begin{aligned} I(y) = I(x^{-1/3}) &= 3^{-2/3} x^{1/3} \int_0^\infty e^{(3v)^{1/3} - xv} v^{-2/3} dv \\ &\sim 3^{-2/3} x^{1/3} \int_0^\infty \sum_{k=0}^\infty \frac{(3v)^{k/3}}{k\,!} e^{-xv} dv \\ &\sim \sum_{k=0}^\infty 3^{(k-2)/3} \frac{\Gamma\left(\frac{k+1}{3}\right)}{k\,!} y^k \end{aligned}$$

Here we have used the symbol $\sim$ to designate an asymptotic expansion that may not converge. This series converges for all finite y, and is equal to $I(y)$. (Why?)

Example 6.1.6

$$I(x) = \text{P.V.} \int_0^\infty \frac{3^{-2t}}{1-t^2} dt$$

Watson's lemma does not apply to this integral (since the proof of Watson's lemma does not take into account the possibility that the integral may be a principal-value integral). If we ignore the P.V., expand the denominator in powers of t and integrate term by term, we obtain

$$I(x) \sim \sum_{k=0}^\infty \frac{(2k)\,!}{x^{2k+1}} . \tag{6.1.18}$$

It can be shown that

$$I(x) = \int_0^\infty \frac{\sin xt}{1+t^2} dt .$$

If we perform repeated integration by parts in this last integral, then we do indeed obtain the asymptotic expansion in (6.1.18).

Example 6.1.7 If we apply integration by parts to the integral

$$I(x) = \int_0^\infty \frac{\cos xt}{1+t^2}\,dt$$

we obtain the asymptotic series

$$I(x) \sim \frac{0}{x} + \frac{0}{x^2} + \frac{0}{x^3}\,.$$

Indeed it may be shown that

$$I = \pi\, e^{-x}$$

which shows that, as $x \to \infty$, $I(x)$ converges to zero faster than any negative power of x. Therefore it does not have an asymptotic expansion in negative powers of x.

Example 6.1.8 Numerical integration can sometimes replace series expansions. For example, if we apply Gauss-Laguerre quadrature to the integral

$$\int_0^\infty e^{-xt} f(t)dt = \frac{1}{x}\int_0^\infty e^{-t} f\left(\frac{t}{x}\right) dt$$

we get

$$\int_0^\infty e^{-xt} f(t)dt = \frac{1}{x}\sum_{k=1}^{n} w_k f\left(\frac{t_k}{2}\right) + \frac{(n\,!)^2 f^{(2n)}(\xi)}{(2n)\,!\,x^{2n+1}} \tag{6.1.19}$$

for some $\xi = \xi(x) \in (0,\infty)$. The error term in (6.1.19) is obtained using the formula (5.6.27), we get

$$\left(\frac{d}{dt}\right)^{2n} f\left(\frac{t}{x}\right) = \frac{1}{x^{2n}}\left(\frac{d}{d(t/x)}\right)^{2n} f\left(\frac{t}{x}\right)\,.$$

Example 6.1.9 Watson's lemma can be readily adapted to integrals over a finite interval. In the case of the integral

$$I(x) = \int_0^T e^{-xt} t^m f(t)dt$$

we have two choices: (a) we can set $f(t) = 0$ for $t > T$ and replace $\int_0^T$ by $\int_0^\infty$, or (b) if $f(t)$ satisfies the conditions of Watson's lemma on the interval $(0,\infty)$, we can write

$$\int_0^T = \int_0^\infty - \int_T^\infty$$

and under the conditions of the lemma, in particular (6.1.15)

$$\begin{aligned}
\left|\int_T^\infty e^{-xt} t^m f(t)dt\right| &< C\int_T^\infty e^{-(x-b)t} t^m dt \\
&= C\int_0^\infty e^{-(x-b)(T+\tau)} (T+\tau)^m d\tau \\
&\le C\,e^{-(x-b)T}\int_0^\infty e^{-(x-b)\tau}[(2T)^m + (2\tau)^m]d\tau \\
&= C\,e^{-(x-b)T}\left[\frac{(2T)^m}{x-b} + \frac{2^m\Gamma(m+1)}{(x-b)^{m+1}}\right]
\end{aligned}$$

Problems

6.1.1 Perform one more integration by parts in (6.1.10) to see that the method of integration by parts also leads to the second term on the extreme right of (6.1.13).

6.1.2 Show that

$$(1+x)^\alpha = \sum_{j=0}^{\infty} \frac{\alpha_j}{j\,!}\, x^j$$

where $\alpha_0 \equiv 1$, $\alpha_j = \alpha(\alpha-1)\cdots(\alpha-j+1)$. Note that if $\alpha = -1/2$ then

$$\alpha_j = \frac{(2j-1)\,!\,!}{2^j}.$$

6.1.3 Show that

$$I(k) \equiv \int_0^\infty v^k e^{-r}\, dr = k\,!\,.$$

Hint: Apply integration by parts to $I(k)$.

6.1.4 Do the term by term integration indicated in Example 5. Problem 6.2 with x replaced by $-x$ and $\alpha = -1$ is useful.

6.1.5 Perform two integrations by parts in

$$I(x) = \int_0^\infty \frac{\sin(xt)}{t^2+1}\, dt$$

to verify the first two terms of (6.1.18). In each step take the differentiated term to be the trigonometric function.

6.2 Definitions and Notations

In the previous section we have been using asymptotic series in powers of x^{-1} or x^{-2}. In this section we introduce the standard concept of a more general asymptotic series.

Definition 6.2.1 *Let f and ϕ be two functions defined on a domain R. We write*

$$f = 0(\phi)$$

if there exists a positive constant A such that

$$|f(x)| \leq A|\phi(x)|$$

for all $x \in R$. We also write $f = 0(\phi)$ as $x \to x_0$ if there is an open subset R_0 of R containing a point x_0 and a constant A, such that $|f(x)| \leq A|\phi(x)|$ for all $x \in R_0$.

Definition 6.2.2 *We write*

$$f = o(\phi)$$

as $x \to x_0$ if for any $\epsilon > 0$ there exists an open subset R_0 of R such that

$$|f(x)| \leq \epsilon|\phi(x)|$$

for all $x \in R_0$.

If $\phi(x) \neq 0$ on R, $x \neq x_0$, then $f = 0(\phi)$ means that $|f/\phi|$ is bounded, and $f = o(\phi)$ as $x \to x_0$ means that $|f/\phi|$ tends to zero as $x \to x_0$.

Definition 6.2.3 *A sequence of functions $\{\phi_k(x)\}_{k=1}^{\infty}$ is called an asymptotic sequence if there exists a neighborhood of x_0 where $\phi_k(x) \neq 0$ when $x \neq x_0$, $k = 0, 1, 2, \ldots$, and where*

$$\phi_{k+1} = 0(\phi_k)$$

as $x \to x_0$, $k = 0, 1, 2, \ldots$.

For example, the sequence $\{x^{-k}\}_{k=1}^{\infty}$ is an asymptotic sequence, with $x_0 = \infty$, and the neighborhood of x_0 being $\{x : x > 0\}$; similarly if $x_0 = -\infty$; the neighborhood of x_0 being $\{x : x < 0\}$.

Definition 6.2.4 *Let $\{\phi_k\}$ be an asymptotic sequence. The series $\sum_{k=0}^{\infty} a_k\phi_k$ is said to be an asymptotic expansion of f as $x \to x_0$ and we write*

$$f \sim \sum_{k=0}^{\infty} a_k\phi_k \ , \quad x \to x_0$$

if for every integer $n = 1, 2, 3, \dots$,

$$f - \sum_{k=0}^{n-1} a_k \phi_k = 0(\phi_n)$$

as $x \to x_0$.

Example 6.2.5 In the asymptotic expansion of $Ei(x)$ in (6.1.3), the asymptotic sequence has

$$\phi_k(x) = e^{-x} x^{-(k+1)} \ , \quad k = 0, 1, 2, \dots$$

and

$$a_k = (-1)^k \, k\,! \ , \quad k = 0, 1, 2, \dots$$

Example 6.2.6 In the asymptotic expansion of $Erfc(x)$ in (6.1.13) the asymptotic sequence has

$$\phi_k(x) = e^{-x^2} x^{1-2k} \ , \quad k = 0, 1, 2, \dots$$

and

$$\sqrt{\pi}\, a_0 = 1 \ , \ \sqrt{\pi}\, a_k = (-1)^k (2k-1)!! \ , \quad k = 1, 2, 3, \dots .$$

6.3 The Method of Laplace

Integrals of the form

$$I(x) = \int_a^b e^{-xg(t)} f(t) dt \tag{6.3.1}$$

are called Laplace-type integrals, after Laplace, who developed an ingenious method of obtaining asymptotic approximation of these integrals. Let us briefly describe the method of obtaining a Laplace-type approximation of the integral (6.3.1).

Assumption 6.3.1 *We suppose that the function* $\exp[-xg(t)]$ *has a unique maximum, located at a point* τ *in* $[a, b]$. *The following three possibilities occur most frequently (there are others that we omit):*

(i) τ *is an endpoint of* $[a, b]$, *e.g.* $\tau = a$ *and* $g'(\tau) \neq 0$.

(ii) τ *is an endpoint of* $[a, b]$, *e.g.* $\tau = a$ *and* $g'(\tau) = 0$, $g''(\tau) > 0$.

(iii) τ *is an interior point of* $[a, b]$, $g'(\tau) = 0$, $g''(\tau) > 0$.

Theorem 6.3.2 *Let one of the above possibilities be satisfied, let a and b be finite and let f be continuous on $[a,b]$.*

(i) If Assumption 6.3.1-(i) holds, and $\tau = a$ then

$$I(x) = \frac{e^{-xg(a)}}{xg'(a)}\left[f(a) + o(1)\right], \quad x \to \infty \tag{6.3.2}$$

(ii) If Assumption 6.3.1-(ii) holds, and $\tau = a$ then

$$I(x) = \left[\frac{\pi}{2xg''(a)}\right]^{1/2} e^{-xg(a)}[f(a) + o(1)], \quad x \to \infty \tag{6.3.3}$$

(iii) If Assumption 6.3.1-(iii) holds, then

$$I(x) = \left[\frac{2\pi}{xg''(\tau)}\right]^{1/2} e^{-xg(\tau)}[f(\tau) + o(1)], \quad x \to \infty \tag{6.3.4}$$

We do not prove this theorem, since the arguments differ little from those of the proof of the Laplace expansion method that we carry out in the next section. We do however give a simple, intuitive explanation of the method of Laplace, for the special case

$$y(x) = \int_{-a}^{a} e^{-xt^2} f(t)\,dt\,. \tag{6.3.5}$$

The graph of $\exp(-xt^2)$ is shown in Figure 6.3.1.

As x increases, the neighborhood of $t = 0$ that contributes significantly to $y(x)$ in (6.3.5) becomes smaller and smaller.

Thus it makes no appreciable difference whether we replace $y(x)$ in (6.3.5) by
$\int_{-\epsilon}^{+\epsilon} e^{-xt^2} f(t)dt$, where ϵ is small but positive or by $\int_{-\infty}^{\infty} \exp(-xt^2) f(t)dt$ (in this latter case we set $f(t) = 0$ for $t > a$). In the region $-\epsilon \le t \le \epsilon$, $f(t) \cong f(0)$, and

$$y(x) \sim f(0)\int_{-\infty}^{\infty} e^{-xt^2}dt = \left(\frac{\pi}{x}\right)^{1/2} f(0)\,, \quad x \to \infty\,.$$

The following sequence of examples illustrates the uses of Laplace's theorem in each of the various cases ???(i)–(iii) cited above.

Figure 6.3.1: The graph of $exp(-xt^2)$.

Example 6.3.3 Consider the integral

$$I_k(x) = \int_0^\pi \cos(t) e^{-xt^k} dt \ , \quad k = 1 \text{ or } 2 \, .$$

The cases $k = 1$ or 2 are handled by Theorem 6.3.2-(i) or (ii) respectively and yield with $\tau = 0$

$$I_1(x) = \frac{1}{x} [1 + o(1)] \ , \quad x \to \infty \, ,$$

and

$$I_2(x) = \frac{1}{2} \left(\frac{\pi}{x} \right)^{1/2} [1 + o(1)] \ , \quad x \to \infty \, .$$

Example 6.3.4 The integral

$$I(x) = \int_{-a}^a e^{-x \cosh(t)} f(t) dt \ , \quad a > 0$$

is handled by case (iii) with $\tau = 0$. Equation (6.3.4) gives

$$I(x) = \left(\frac{2\pi}{x} \right)^{1/2} e^{-x} [f(0) + o(1)] \ , \quad x \to \infty \, .$$

Example 6.3.5 The integral

$$I(x, y) = \int_{-\infty}^\infty e^{-xt^2} \cos(yt) dt$$

has radically different behavior for large x (versus large y). Here we consider the former case. If $a > 1$ then

$$\begin{aligned} \left| \int_a^\infty e^{-xt^2} \cos yt \, dt \right| & \leq \int_a^\infty e^{-xt^2} dt \\ & \leq \tfrac{1}{2} \int_x^\infty e^{-u} du \\ & = \tfrac{1}{2} e^{-x} \quad . \end{aligned}$$

By Theorem 6.3.2-(iii)

$$\int_{-a}^{a} e^{-xt^2}\cos(yt)dt = \left(\frac{\pi}{x}\right)^{1/2}[1+o(1)]\,, \quad x\to\infty$$

so that

$$I(x,y) = \left(\frac{\pi}{x}\right)^{1/2}[1+o(1)] + \delta(x)$$

where

$$|\delta(x)| \le e^{-x}\,.$$

That is

$$I(x,y) = \left(\frac{\pi}{x}\right)^{1/2}[1+o(1)]\,, \quad x\to\infty\,.$$

It can shown that

$$I(x,y) = \left(\frac{\pi}{x}\right)^{1/2} e^{-y^2/4x}$$

so that the approximation $(\pi/x)^{1/2}$ is quite accurate for small $\left(y^2/(4x)\right)$.

Example 6.3.6 If n is an integer the n^{th} Bessel function may be defined by

$$I_n(x) = \frac{1}{\pi}\int_0^{\pi} e^{x\cos\theta}\cos(n\theta)d\theta\,.$$

It follows from Theorem 6.3.2-(ii) that

$$I_n(x) = \frac{e^x}{\sqrt{2\pi x}}[1+o(1)]\,, \quad x\to\infty\,.$$

Example 6.3.7 In the previous example set $u=\cos\theta$ and write

$$I_n(x) = \frac{1}{\pi}\int_{-1}^{1} e^{xu}\frac{T_n(u)}{(1-u^2)^{1/2}}\,du$$

where $T_n(u) = \cos\left(n\cos^{-1}(u)\right)$ is the n^{th} Chebyshev polynomial of degree n (Chapter 4). Now set $u=1-t$ to get

$$I_n(x) = \frac{1}{\pi}\int_0^{2}\frac{e^{x(1-t)}T_n(1-t)}{t^{1/2}(2-t)^{1/2}}\,dt\,.$$

Using ($a=\tau=0$) Theorem 6.3.2-(i) is not applicable since $f(t)=T_n(1-t)(2-t)^{-1/2}t^{-1/2}$ is not continuous on any subinterval of $[0,2]$ containing zero. However, replacing t by t^2 leads to

$$\begin{aligned} I_n(x) &= \frac{2}{\pi}\int_0^{\sqrt{2}}\frac{e^{x(1-t^2)}T_n(1-t^2)dt}{(2-t^2)^{1/2}} \\ &= \frac{e^x}{\sqrt{2\pi x}}[T_n(1)+o(1)] \\ &= \frac{e^x}{\sqrt{2\pi x}}[1+o(1)] \end{aligned}$$

where Theorem 6.3.2-(ii) has again been used.

Example 6.3.8 The Gamma function is defined by

$$\Gamma(x+1) = \int_0^\infty e^{-u} u^x du$$

and can be turned into a Laplace type integral by the substitution

$$u = x + 1 \, .$$

Thus

$$\begin{aligned}\Gamma(x+1) &= x^{x+1} e^{-x} \int_{-1}^\infty e^{-x[t - \ell n(t+1)]} dt \\ &= (2\pi x)^{1/2} \left(\frac{x}{e}\right)^x [1 + o(1)]\end{aligned}$$

where Theorem 6.3.2-(iii) with $\tau = 0$ was used.

In Problem 6.3.3 the complete asymptotic expansion of $I_0(x)$ was found by converting the integral to one for which Watson's lemma applies. This procedure is more general than that problem may have indicated.

Theorem 6.3.9 *In the integral*

$$I(x) = \int_a^b e^{-xg(t)} f(t) dt \tag{6.3.6}$$

let Assumption 6.3.1-(i) be satisfied. Let $g'(t) > 0$ on $[a,b]$ and define $t = h(w)$ through the equation

$$w = g(t) - g(a) \, , \, a \le t \le b \, . \tag{6.3.7}$$

If the function H satisfies

$$H(w) \equiv f(h(w))h'(w) = \sum_{k=0}^\infty \frac{H^{(k)}(0) w^k}{k\,!} \, , \quad |w| \le A \tag{6.3.8}$$

and

$$H(w) \le C\, e^{\beta w} \, , \, |w| > A \tag{6.3.9}$$

where

$$A = g(b) - g(a) \tag{6.3.10}$$

then

$$I(x) \sim e^{-xg(a)} \sum_{k=0}^\infty \frac{H^{(k)}(0)}{k\,!\, x^{k+1}} \, , \quad x \to \infty \, . \tag{6.3.11}$$

Theorem 6.3.10 *If Assumption 6.3.1-(ii) is satisfied and* $g'(t) > 0$ *on* $(a, b]$*, set*

$$w = [g(t) - g(a)]^{1/2} \tag{6.3.12}$$

and suppose that H *satisfies (6.3.8) and*

$$H(w) \leq C\, e^{\beta w^2}\ ,\ |w| > A \tag{6.3.13}$$

where

$$A = [g(b) - g(a)]^{1/2}\,. \tag{6.3.14}$$

In this case (6.3.6) satisfies

$$I(x) \sim \frac{e^{-xg(a)}}{2} \sum_{k=0}^{\infty} \frac{H^{(k)}(0)\Gamma(k+1/2)}{k\,!\,x^{(k+1/2)}}\ ,\quad x \to \infty\,. \tag{6.3.15}$$

Theorem 6.3.11 *If Assumption 6.3.1-(iii) is satisfied and* $g'(t) < 0$ *on* $a < t < \tau$*, set*

$$w = \begin{cases} -[g(t) - g(\tau)]^{1/2} & ,\ a \leq t \leq \tau \\ [g(t) - g(\tau)]^{1/2} & ,\ \tau \leq t \leq b \end{cases}\,. \tag{6.3.16}$$

Assume that H *satisfies (6.3.8) and (6.3.13) with*

$$A = \min\{[g(b) - g(\tau)]^{1/2}\ ,\ [g(a) - g(\tau)]^{1/2}\}\,. \tag{6.3.17}$$

In this case (6.3.6) satisfies

$$I(x) \sim e^{-xg(\tau)} \sum_{k=0}^{\infty} \frac{H^{(2k)}(0)\Gamma(k+1/2)}{(2k)\,!\,x^{k+1/2}}\ ,\quad x \to \infty\,. \tag{6.3.18}$$

Proof: Since the proofs of each of the three theorems are similar we only prove the last one.

If the conditions of Theorem 6.3.11 are satisfied then

$$\begin{aligned} J(x) &\equiv \int_a^b e^{-x\left(g(t)-g(\tau)\right)} f(t)dt + \epsilon \\ &= \int_{-A}^{A} e^{-xw^2} H(w)dw + \epsilon \\ &= e^{-xg(\tau)} I(x) + \epsilon \end{aligned} \tag{6.3.19}$$

where

$$
\begin{aligned}
|\epsilon| &\le 2C\int_A^\infty e^{-(x-\beta)w^2}dw \\
&= 2C\int_{(x-\beta)A^2}^\infty \frac{e^{-y}}{(x-\beta)^{1/2}}\,y^{-1/2}dy \qquad (6.3.20)\\
&\le \frac{C\,e^{-(x-\beta)A^2}}{(x-\beta)^{1/2}}\,,\ (x-\beta)A^2\ge 1\,.
\end{aligned}
$$

Thus $\epsilon = o(x^{-m})$ as $m\to\infty$ for all $m>0$.

Using (6.3.8) in (6.3.19) gives

$$
\begin{aligned}
\int_{-A}^{A} e^{-xw^2}H(w)dw &= \int_{-A}^{A} e^{-xw^2}\sum_{k=0}^\infty \frac{H^{(k)}(0)}{k\,!}\,w^k dw \\
&= \int_{-A}^{A} e^{-xw^2}\sum_{k=0}^\infty \frac{H^{(2k)}(0)}{(2k)\,!}\,w^{2k}dw\,.
\end{aligned}
\qquad (6.3.21)
$$

Putting $w^2=v$ in (6.3.21) gives

$$
\begin{aligned}
\int_{-A}^{A} e^{-xw^2}H(w)dw &= \int_0^A e^{-xv}\sum_{k=0}^\infty \frac{H^{2k}(0)}{(2k)\,!}\,v^{k-1/2}dv \\
&\sim \sum_{k=0}^\infty \frac{H^{(2k)}(0)}{(2k)\,!}\,\frac{\Gamma(k+1/2)}{x^{k+1/2}}
\end{aligned}
\qquad (6.3.22)
$$

by Watson's lemma. Combining (6.3.20) and (6.3.22) in (6.3.19) gives (6.3.18).

Example 6.3.12 From (6.3.7)

$$\frac{dw}{dt} = g'(t)\Big|_{t=a} = g'(a)$$

which combined with the change of variable $t=h(w)$ shows that the leading term of (6.3.11) is the same as (6.3.2).

Example 6.3.13 Integrals of the form

$$I(n) = \int_a^b g(t)^n f(t)dt$$

can sometimes be brought into the form of a Laplace integral via

$$-g(t) = \ell n\big(h(t)\big)\,.$$

Thus if h has exactly one maximum at $\tau \in [a,b]$ then

$$\begin{aligned} I(n) &= \int_a^b e^{-ng(t)} f(t)dt \\ &= \left[\frac{2\pi}{ng''(\tau)}\right]^{1/2} e^{-ng(\tau)}[f(\tau)+o(1)] \qquad (6.3.23) \\ &= \left[\frac{2\pi}{-nh''(\tau)}\right]^{1/2} [h(\tau)]^{n+1/2}[f(\tau)+o(1)]\ , \quad n\to\infty\,. \end{aligned}$$

Example 6.3.14 The Legendre polynomial $P_n(x)$ can be defined by

$$P_n(x) = \frac{1}{\pi}\int_0^\pi [x+(x^2-1)^{1/2}\cos t]^n dt\ , \quad x>1\,.$$

Use Example 8 with $f(t)=1$ and $h(t)=x+(x^2-1)^{1/2}\cos t$. Now h has a maximum at $t=0$ and $h''(0)=-(x^2-1)^{1/2}$ so from (6.3.23)

$$P_n(x) = \frac{1}{2\pi}\left[\frac{2\pi}{n(x^2-1)^{1/2}}\right]^{1/2} [x+(x^2-1)^{1/2}]^{n+1/2}[1+o(1)]\ , \quad n\to\infty\,.$$

Example 6.3.15 The function $h(w)$ in the previous theorems may be impossible to obtain in closed form. In this case the Gauss-Hermite quadrature proves helpful. With Assumption 6.3.1-(ii) and (6.3.6) we can write

$$I(x) = e^{-xg(\tau)}\int_a^b e^{-xg''(\tau)(t-\tau)^2/2} H(x,t)dt \qquad (6.3.24)$$

where

$$H(x,t) = \exp\{-x[g(t)-g(\tau)-(1/2)(t-\tau)^2 g''(\tau)]\}f(t)\,. \qquad (6.3.25)$$

In (6.3.24) set

$$t = \tau + \left[\frac{2}{xg''(\tau)}\right]^{1/2} y \equiv \tau + Cy$$

to find

$$\begin{aligned} I(x) &= C\,e^{-xg(\tau)}\int_{-\infty}^{\infty} e^{-y^2} H(x,\,\tau+Cy)dy \\ &\sim C\,e^{-xg(\tau)}\left[\sum_{k=1}^m w_k H(x,\,\tau+Cy_k)+\epsilon\right] \end{aligned} \qquad (6.3.26)$$

where the y_k and w_k are the nodes and weights of Gauss-Hermite quadrature. From (5.6.28), the error in (6.3.26) is

$$\epsilon = \frac{\sqrt{\pi}\ m\,!}{2^{2m}(2m)\,!}\left(\frac{\partial}{\partial y}\right)^{2m} H(x,\,\tau+Cy) \quad . \qquad (6.3.27)$$

We need only consider the exponential component of $H(x,y)$ in estimating the error ϵ and only the first term in the expansion

$$\begin{aligned} \exp\{ & -x[g(t)-g(\tau)-(1/2)(t-\tau)^2 g''(\tau)]\} \\ & = \exp\{-x[g_3 c^3 y^3 + g_4 c^4 y^4 + \cdots]\}\,. \end{aligned} \tag{6.3.28}$$

However we have

$$\begin{aligned} \left(\frac{d}{dy}\right)^{2m} [\exp(\alpha y^3)] &= \alpha^{2m/3}\left(\frac{d}{dw}\right)^{2m} [\exp(w^3)] \\ &= 0(\alpha^{2m/3}) \\ &= 0(x^{-m/3}) \end{aligned}$$

so it follows that

$$\epsilon = 0(x^{-m/3})\,.$$

Problems

6.3.1 Write the integral in Example 2 as

$$I(x) = 2\int_0^a e^{-x\cosh(t)} f(t)dt\,.$$

Compare the result of Example 2 with the application of Theorem 6.3.2-(ii) to $I(x)$.

6.3.2 Apply Theorem 6.3.2 to the integral

$$I(x) = \int_{-\pi}^{\pi} e^{x\cos(t)} f(t)dt\,.$$

6.3.3 Use Watson's lemma to obtain the asymptotic expansion of

$$\begin{aligned} I_0(x) &= \frac{1}{\pi}\int_0^2 \frac{e^{x(1-t)}T_0(1-t)}{(2-t)^{1/2}t^{1/2}}\,dt \\ &= \frac{1}{\pi}\int_0^2 e^{-xt}t^{-1/2}\sum_{n=0}^{\infty}\frac{(1/2)_n t^n}{2^n n\,!}\,dt \\ &\sim \frac{e^x}{\sqrt{2\pi x}}\sum_{n=0}^{\infty}\frac{[(1/2)_n]^2}{2^n n\,!\,x^n}\,. \end{aligned}$$

See Example 8, Section 6.1.

6.3.4 Apply Theorem 6.3.2-(ii) to

$$I(x) = \int_{\pi}^{\infty} \cos(t) e^{-xt^2} dt\,.$$

6.3.5 The leading term in the result of Example 6 is called Stirling's formula, i.e.

$$n\,! \sim \sqrt{2\pi n}\,(n/e)^n\,.$$

It has been stated a number of times that the series

$$\sum_{n=0}^{\infty} \frac{n\,!}{x^{n+1}}$$

diverges for all positive x. Show this using Stirling's formula.

6.4 Darboux's Method

Darboux's method is particularly valuable in numerical work since it permits estimates of functions defined by a generating function and is thus of value for example in obtaining asymptotic estimates of orthogonal polynomials.

The concept the method depends on is best illustrated by a simple example. Consider the Taylor series development

$$\begin{aligned} \frac{1}{1-ax}\,e^x &= (1+ax+a^2x^2+\cdots)\left(1+x+\frac{x^2}{2\,!}+\cdots\right) \\ &= 1+(a+1)x+\left(a^2+\frac{a}{1\,!}+\frac{1}{2}\right)x^2+\left(a^3+\frac{a^2}{1\,!}+\frac{a}{2\,!}+\frac{1}{3\,!}\right)x^3+\cdots \end{aligned} \tag{6.4.1}$$

by straightforward multiplication of series. The coefficient of x^n can be written

$$a^n\left(1+\frac{1}{1\,!\,a}+\frac{1}{2\,!\,a^2}+\cdots+\frac{1}{n\,!\,a^n}\right)$$

so that the expression in parenthesis is the n^{th} partial sum of $\exp(1/n)$. Hence the coefficient of x^n in (6.4.1) approaches $a^n \exp(1/n)$. In particular, for $a = 1$

$$F(x) = \frac{1}{1-x}\,e^x = \sum_{n=0}^{\infty} F_n x^n \tag{6.4.2}$$

where

$$F_n \to e \,\text{as}\, n \to \infty\,.$$

Hence, if $F(x)$ is the product of a function with a simple singularity at $x = 1$ and a function that has no singularity at the origin we surmise that the coefficients, F_n, of the Taylor series for $F(x)$ are proportional to $F(1)$ ($= e$ in (6.4.2)) are $n \to \infty$.

We want to obtain an asymptotic estimate of the Taylor coefficients F_n where

$$F(x) = \frac{1}{(1-x)^\alpha}\, f(x) = \sum_{n=0}^{\infty} F_n x^n\,. \tag{6.4.3}$$

If

$$f(x) = \sum_{n=0}^{\infty} f_n x^n\,, \quad |x| < 1+\epsilon \tag{6.4.4}$$

where $f_0 \neq 0$ and $\epsilon > 0$ then (6.4.3) has a singularity at $x = 1$ as long as α is not a negative integer. We also assume that the Taylor series of f converges at $x = 1$, i.e. for $\mu > 0$

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(1)}{n\,!}\,(x-1)^n\,, \quad |x-1| < \mu\,. \tag{6.4.5}$$

Theorem 6.4.1 (Darboux) *Suppose that f in (6.4.3) satisfies (6.4.4) and (6.4.5). Then the coefficients F_n in (6.4.3) satisfy*

$$F_n = \frac{n^{\alpha-1}}{\Gamma(\alpha)}\, f(1)[1 + 0(1/n)]\,, \quad n \to \infty\,. \tag{6.4.6}$$

Proof: Using (6.4.5) in (6.4.3) we have

$$F(x) = \frac{1}{(1-x)^\alpha}\, f(1) - (1-x)^{1-\alpha} f'(1) + \frac{(1-x)^{2-\alpha}}{2\,!}\, f''(1) - \cdots\,.$$

Expanding the factors $(1-x)^{n-\alpha}$ in powers of x ($|x| < 1$) we obtain

$$F(x) = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} \left\{ \frac{(-1)^k f^{(k)}(1)}{k\,!}\,(\alpha - k + n - 1)_{(n)} \right\} x^n\,. \tag{6.4.7}$$

The exchange of order of summation is permitted since the series converges in the common region $\{x : |x| < 1 \text{ and } |x-1| < \mu\}$. Comparing (6.4.3) and (6.4.7) coupled with the uniqueness of power series expansions gives

$$F_n = \sum_{k=0}^{\infty} \frac{(-1)^k f^{(k)}(1)}{k\,!} \, (\alpha - k + n - 1)_{(n)} \, . \tag{6.4.8}$$

Since (Problem 6.4.3)

$$\frac{(\alpha - k - \ell + n - 1)_{(n)}}{(\alpha - k + n - 1)_{(n)}} = 0\left(\frac{1}{n^{\ell}}\right) \quad , \quad \ell = 1, 2, \ldots \; , \quad n \to \infty$$

only the first term in the series (6.4.8) is significant, i.e.

$$\begin{aligned} F_n &= (\alpha + n - 1)_{(n)} f(1)[1 + 0(1/n)] \; , \quad n \to \infty \\ &= \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)} \, f(1)[1 + 0(1/n)] \; , \quad n \to \infty \\ &= \frac{n^{\alpha - 1}}{\Gamma(\alpha)} \, f(1)[1 + 0(1/n)] \; , \quad n \to \infty \end{aligned} \tag{6.4.9}$$

where the last equality uses Stirling's formula.

Example 6.4.2 The Bernouli numbers B_n are defined by

$$\frac{t}{e^t - 1} = \sum_{n=0}^{\infty} B_n \frac{t^n}{n\,!} \, . \tag{6.4.10}$$

Recall Problem ??5.4.2 with $x = 0$. Adding $t/2$ to the left-hand side of (6.4.10) gives

$$\frac{t}{e^t - 1} + \frac{t}{2} = \frac{t}{2}\left(\frac{e^t + 1}{e^t - 1}\right)$$

which is an even function of t. Hence

$$B_1 = -1/2 \; , \; B_{2n+1} = 0 \; , \; n = 1, 2, \ldots$$

and

$$\frac{t}{2}\left(\frac{e^t + 1}{e^t - 1}\right) = \sum_{n=0}^{\infty} B_{2n} \frac{t^{2n}}{(2n)\,!} \, . \tag{6.4.11}$$

The left-hand side of (6.4.11) has a singularity at $t = 2\pi i$ so by writing $t = 2\pi i x$ in (6.4.11) the singularity is transformed to $x = 1$;

$$\begin{aligned} \pi i x \left(\frac{e^{2\pi i x} + 1}{e^{2\pi i x} - 1}\right) &= \sum_{n=0}^{\infty} (-1)^n B_{2n} \, \frac{(2\pi)^{2n}}{2n\,!} \, x^{2n} \\ &= \frac{\pi x \cos(\pi x)}{\sin(\pi x)} \, . \end{aligned} \tag{6.4.12}$$

Now $x/\bigl(\sin(\pi x)\bigr)$ is singular at $x = \pm 1$ and we eliminate these singularities by defining

$$f(x) = \pi x(x^2 - 1)\,\frac{\cos(\pi x)}{\sin(\pi x)}\,.$$

Now f is an even function of x $f(1) = f(-1) = 2$ so the generating function (6.4.12) for the B_{2n} can be written

$$\frac{f(x)}{1 - x^2} = \sum_{n=0}^{\infty} (-1)^n\,\frac{B_{2n}(2\pi)^{2n}}{(2n)\,!}\,x^{2n}\,. \tag{6.4.13}$$

Replacing x^2 by y in (6.4.13) and using Darboux's theorem gives

$$(-1)^{n+1} B_{2n}\frac{(2\pi)^{2n}}{(2n)\,!} = f(1)[1 + 0(1/n)]\ ,\quad n \to \infty\,.$$

That is

$$B_{2n} = (-1)^{n+1} 2\frac{(2n)\,!}{(2\pi)^{2n}}\,[1 + 0(1/n)]\ ,\quad n \to \infty\,.$$

Example 6.4.3 The Chebyshev polynomials are generated by

$$f(t) = \frac{1 - tx}{1 - 2tx + t^2} = \sum_{n=0}^{\infty} T_n(x) t^n\,.$$

Now $1 - 2tx + t^2$ vanishes at

$$t = x \pm (x^2 - 1)^{1/2}\ ,\quad x > 1$$

so if we put $\beta = x + (x^2 - 1)^{1/2} > 1$ then

$$1/\beta = x - (x^2 - 1)^{1/2}$$

and

$$x = \frac{1}{2}\,(\beta + 1/\beta)\,.$$

Hence

$$\begin{aligned} f(t) &= \frac{1 - 1/2t(\beta + \beta^{-1})}{\beta - \beta^{-1}}\left[\frac{1}{\beta^{-1} - t} - \frac{1}{\beta - t}\right] \\ &\equiv F_1(t) + F_2(t) \end{aligned} \tag{6.4.14}$$

where

$$F_1(t) = \left(\frac{\beta^2 - 1/2\beta t(\beta^2 + 1)}{\beta^2 - 1}\right)\left(\frac{1}{1 - \beta t}\right) \tag{6.4.15}$$

and

$$F_2(t) = -\left(\frac{\beta - 1/2t(\beta^2 + 1)}{\beta(\beta^2 - 1)}\right)\left(\frac{1}{1 - \beta^{-1} t}\right)\,. \tag{6.4.16}$$

In Problem 6.4.4 you are asked to show that if

$$F(t) = \frac{1}{(1-\gamma t)^\alpha} f(t) = \sum_{n=0}^{\infty} F_n t^n \tag{6.4.17}$$

then the conclusion of Darboux's theorem (6.4.6) reads

$$F_n = \frac{\gamma^n n^{\alpha-1}}{\Gamma(\alpha)} f(\gamma^{-1})[1 + 0(1/n)]\,. \tag{6.4.18}$$

Putting $F_j(t) = \sum_{n=0}^{\infty} F_{jn} t^n$ $(j = 1, 2)$ we have by (6.4.18), using $\gamma = \beta$ in F_1 and $\gamma = \beta^{-1}$ in F_2,

$$F_{1n} = \frac{1}{2}\beta^n[1 + 0(1/n)]\ , \quad n \to \infty$$

and

$$F_{2n} = \frac{1}{2}\beta^{-n}[1 + 0(1/n)]\ , \quad n \to \infty\,.$$

Hence

$$\begin{aligned} T_n(x) &= F_{1n} + F_{2n} \\ &= \frac{1}{2}(\beta^n + \beta^{-n})[1 + 0(1/n)] \\ &= \frac{1}{2}\beta^n[1 + 0(1/n)] \\ &= \frac{1}{2}[x + (x^2-1)^{1/2}]^n[1 + 0(1/n)]\,. \end{aligned}$$

Example 6.4.4 The Legendre polynomials $P_n(\cos\theta)$, $0 < \theta < \pi$, can be defined by the generating function

$$F(t) = \frac{1}{(1 - 2t\cos\theta + t^2)^{1/2}} = \sum_{n=0}^{\infty} P_n(\cos\theta) t^n\,.$$

Rewriting

$$F(t) = \frac{1}{(1-te^{i\theta})^{1/2}(1-te^{-i\theta})^{1/2}}$$

then P_n can be asymptotically estimated as $n \to \infty$ using (6.4.17) and (6.4.18). In the notation of (6.4.17) with $\beta = e^{i\theta}$ and $\alpha = 1/2$ we have

$$\begin{aligned} P_n(\cos\theta) &= \tfrac{n^{-1/2}}{\Gamma(1/2)}\left[\tfrac{e^{in\theta}}{(1-e^{2i\theta})^{1/2}} + \tfrac{e^{-in\theta}}{(1-e^{-2i\theta})^{1/2}}\right][1 + 0(1/n)] \\ &= \tfrac{1}{(\pi n)^{1/2}}[u + \hat{u}][1 + 0(1/n)] \end{aligned} \tag{6.4.19}$$

where, with $i^{1/2} = e^{\pi/4i}$,

$$u = \frac{e^{in\theta}}{(1-e^{2i\theta})^{1/2}} = \frac{e^{in\theta + i\theta/2}}{\left[\left(\frac{e^{i\theta}-e^{-i\theta}}{2i}\right)2i\right]^{1/2}} = \frac{e^{i[(n+1/2)\theta - \pi/4]}}{\sqrt{2}\,\sin(\theta)} \tag{6.4.20}$$

and $\hat{u}$ is obtained by replacing i by $-i$ in (6.4.20). Hence

$$\begin{aligned} P_n(\cos\theta) &= \frac{1}{(\pi n)^{1/2}}\left\{\frac{e^{i[(n+1/2)\theta-\pi/4]}}{\sqrt{2}\sin^{1/2}\theta}+\frac{e^{-i[(n+1/2)\theta-\pi/4]}}{\sqrt{2}\sin^{1/2}\theta}\right\}[1+0(1/n)] \\ &= \left(\frac{2}{\pi n\sin\theta}\right)^{1/2}\cos[(n+1/2)\theta-\pi/4][1+0(1/n)] \\ &= \left(\frac{2}{\pi n\sin\theta}\right)^{1/2}\cos[(n+1/2)\theta-\pi/4]+0\left(\frac{1}{n^{3/2}}\right). \end{aligned}$$

Problems

6.4.1 Using the series development

$$\frac{1}{(1-ax)^2} = 1+2ax+3a^2x^2+4a^3x^3+\cdots \tag{6.4.21}$$

find an estimate of F_n in

$$F(x) = \frac{1}{(1-ax)^2}\,e^x = \sum_{n=0}^{\infty} F_n x^n$$

by the procedure outlined in (6.4.1). What is your result for $a=1$? Check your result via Darboux's theorem.

6.4.2 Obtain the series in (6.4.21). Hint:

$$\frac{d}{dx}\left(\frac{1}{1-ax}\right) = \frac{a}{(1-ax)^2} \qquad \text{and (6.4.1).}$$

6.4.3 Recall that

$$(\gamma)_{(n)} = \gamma(\gamma+1)(\gamma+2)\cdots(\gamma+n-1) = \frac{\Gamma(\gamma+n)}{\Gamma(\gamma)}.$$

Show that

$$\frac{(\gamma-\ell)_{(n)}}{(\gamma)_{(n)}} = 0\left(\frac{1}{n^\ell}\right), \quad \ell=1,2,\ n\to\infty.$$

6.4.4 Obtain from (6.4.17) the formula (6.4.18). Put $\gamma t = x$ in (6.4.17) so that

$$\frac{1}{(1-\gamma t)^\alpha}\,f(t) = \frac{1}{(1-x)^\alpha}\,f(x/\gamma) = \sum_{n=0}^{\infty} F_n x^n.$$

Now consider (6.4.3) and (6.4.6).

6.5 Asymptotic Methods in Differential Equations

This section is concerned with the approximation of the solution of the differential equation

$$\frac{d^2y}{dx^2} + p(x)\,\frac{dy}{dx} + q(x)y = 0 \tag{6.5.1}$$

for large x. Two fairly typical equations are

$$y'' + \left(2 + \frac{1}{x}\right) y' + \frac{y}{x} = 0 \tag{6.5.2}$$

and

$$y'' + \left(1 + \frac{1}{x^2}\right) y = 0 \tag{6.5.3}$$

For large x, we may suspect that the solutions of these equations may, for a first approximation, have the same form as the solutions of

$$\begin{aligned} y'' + 2y' &= 0 \quad (y = A + Be^{-2x}) \\ y'' + y &= 0 \quad (y = A\sin x + B\cos x) \end{aligned}$$

A very simple method for the solution of such equations has been known for more than a century and the solutions so obtained in descending powers of x are called Stokes-type solutions. If $p(x)$ and $q(x)$ in (6.5.1) are bounded, $x \to \infty$, the equation can be written as

$$y'' + \left(a_0 + \frac{a_1}{x} + \cdots\right) y' + \left(b_0 + \frac{b_1}{x} + \cdots\right) y = 0\,. \tag{6.5.4}$$

Assuming that the solutions of the differential equation have the form

$$\begin{aligned} y_1(x) &= e^{\lambda_1 x} x^{\sigma_1} \left[A_0 + \tfrac{A_1}{x} + \tfrac{A_2}{x^2} + \cdots\right] \\ y_2(x) &= e^{\lambda_2 x} x^{\sigma_2} \left[B_0 + \tfrac{B_1}{x} + \tfrac{B_2}{x^2} + \cdots\right] \end{aligned} \tag{6.5.5}$$

the parameters λ_1, λ_2 and σ_1, σ_2 are determined by

$$\begin{aligned} \lambda_1 + \lambda_2 &= -a_0 \qquad \lambda_1\lambda_2 = b_0 \\ \sigma_1 + \sigma_2 &= -a_1 \qquad \lambda_1\sigma_2 + \lambda_2\sigma_1 = b_1 \end{aligned} \tag{6.5.6}$$

The first two of these equations imply that λ_1 and λ_2 are different unless $b_0 = 1/4\, a_0^2$; we exclude this case temporarily. Except in the excluded case, σ_1 and σ_2 are determined, and the remaining coefficients A_i and B_i of (6.5.5) can be determined
sequentially by familiar methods.

Sometimes, this method suffices for the problem in hand and it has the advantage that there are no formulas to remember.

Example 6.5.1 The integral

$$y(x) = \frac{e^{-x}}{\pi} \int_{-1}^{1} e^{-xu}(1-u^2)^{-1/2} du = e^{-x} I_0(x)$$

satisfies the differential equation (6.5.2) where $I_0(x)$ is the modified Bessel function that satisfies $xI_0''(x) + I_0'(x) - xI_0(x) = 0$. Using (6.5.6) with $a_0 = 2$, $a_1 = 1$, $b_0 = 0$ and $b_1 = 1$ we find

$$\begin{aligned} \lambda_1 &= 0 \quad \lambda_2 = -2 \\ \sigma_1 &= \sigma_2 = -1/2 \end{aligned}$$

for (6.5.5). Hence, the solutions of (6.5.2) have the form

$$y_1(x) = x^{-1/2}\left[A_0 + \frac{A_1}{x} + \frac{A_2}{x^2} + \cdots\right]$$

and

$$y_2(x) = e^{-2x} x^{-1/2}\left[B_0 + \frac{B_1}{x} + \frac{B_2}{x^2} + \cdots\right].$$

It is easy to obtain the series solution

$$\begin{aligned} y(x) &= A_0 x^{-1/2}\left[1 + \frac{1}{8x} + \frac{9}{128x^2} + \cdots\right] \\ &\quad + B_0 e^{-2x} x^{-1/2}\left[1 - \frac{1}{8x} + \frac{9}{128x^2} - \cdots\right] \end{aligned}$$

The determination of the constants A_0 and B_0 is always the difficult point in asymptotic solutions of differential equations. Usually the leading terms must be obtained by alternative methods.

Example 6.5.2 For equation (6.5.3) the Stokes procedure gives

$$\lambda_1 + \lambda_2 = 0 \qquad \lambda_1\lambda_2 = 1$$

so that $\lambda_1 = i$ and $\lambda_2 = -i$ and

$$\sigma_1 + \sigma_2 = 0 \qquad i(\sigma_2 - \sigma_1) = 0$$

so that $\sigma_1 = \sigma_2 = 0$. From (6.5.5) the solutions of (6.5.3) have the form

$$y_1(x) = e^{ix}\left[A_0 + \frac{A_1}{x} + \frac{A_2}{x^2} + \cdots\right] \tag{6.5.7}$$

and

$$y_2(x) = e^{-ix}\left[B_0 + \frac{B_1}{x} + \frac{B_2}{x^2} + \cdots\right] \tag{6.5.8}$$

Example 6.5.3 The excluded case of equal roots in the analysis of (6.5.4) ($\lambda_1 = \lambda_2$) presents no grave difficulties. In this case we the factor $\exp(\lambda x)$ from the solution and assume that a_0 and b_0 are zero in (6.5.4). The solution is then of the form

$$y_{\pm}(x) = e^{\pm\mu x^{1/2}} x^{\sigma} \left(A_0 + \frac{A_1}{x} + \frac{A_2}{x^2} + \cdots \right) \tag{6.5.9}$$

where

$$2(\sigma - 1/4) = -a_1 \qquad \frac{1}{4}\mu^2 = b_1 \neq 0\,. \tag{6.5.10}$$

For a number of reasons, mathematical interest has turned towards the study of differential equations simpler in form than (6.5.1). The first of these reasons is that a change of dependent variable yields a differential equation with no dy/dx term. If we write

$$y = u \exp\left[-\frac{1}{2} \int^{x} p(t)dt \right]\,, \tag{6.5.11}$$

then (6.5.1) becomes

$$\frac{d^2u}{dx^2} + \left(q - \frac{1}{2}p' - \frac{1}{4}p^2 \right) u = 0\,. \tag{6.5.12}$$

Hence, we can without loss of generality suppose that the middle term of (6.5.1) absent and in the remainder of this section we are concerned with equations of the form

$$y'' + g(x)y = 0\,. \tag{6.5.13}$$

A second reason for concentration on equations of the type (6.5.13) rather than (6.5.1) is that error analysis of (6.5.13) is much simpler; another reason is that excellent methods are available for (6.5.13).

Even for (6.5.13), it is necessary to imply a simplifying condition, in order to reach a tractable analysis: *In the remainder of this chapter, it is assumed that $g(x)$ is one-signed on the interval where the solution is to be determined.* When g has one sign in one part of the interval and a different sign in another part or when $g(x)$ oscillates between positive and negative values on the interval, the analysis is considerably more difficult and is beyond the scope of this book.

Problems

6.5.1 Show that $y(x) = B_0 + A_0 e^{-2x}$ is the solution of $y'' + 2y' = 0$ as follows: With $p(x) = 2$ we see that $a_0 = 2$ and $a_j = 0$ for all $j \geq 1$. Similarly

$b_j = 0$ $(j \geq 0)$ since $g(x) = 0$. The relations in (6.5.5) then give $\lambda_1 = -2$, $\lambda_2 = 0$, $\sigma_1 = \sigma_2 = 0$ so that

$$y_1(x) = e^{-2x}\left(A_0 + \frac{A_1}{x} + \cdots\right)$$

and

$$y_2(x) = B_0 + \frac{B_1}{x} + \cdots .$$

Substitute each of these in the differential equation to find $A_j = 0$ $(j \geq 1)$ and $B_j = 0$ $(j \geq 1)$.

6.5.2 Find the first few A_j in (6.5.7) as follows: Direct differentiation of $y_1(x)$ gives

$$y_1''(x) + y_1(x) = e^{ix}\left[2i\left(\frac{-A+1}{x^2} - \frac{2A_2}{x^3} - \cdots\right) + \left(\frac{2A_1}{x^3} + \frac{3\,!\,A_2}{x^4} + \cdots\right)\right]$$

and

$$\frac{1}{x^2}\,y_1(x) = e^{ix}\left(\frac{A_0}{x^2} + \frac{A_1}{x^3} + \cdots\right) .$$

Adding these two lines and equating coefficients of like powers gives the A_j $(j \geq 1)$ in terms of A_0 $(A_1 = A_0/2i,\ A_2 = 3/8\,A_0)$. Conclude that

$$y_1(x) = A_0\left\{\sin x\left[1 - \frac{3}{8x^2} + \cdots\right] - \cos x\left[\frac{1}{2x} - \frac{3}{4x^3} + \cdots\right]\right\} .$$

6.5.3 For the differential equation in (6.5.2) show that the equation (6.5.12) takes the form

$$u''(x) + \left[-1 + \frac{1}{4x^2}\right]u(x) = 0\,.$$

As in the line following (6.5.3) we expect (for large x) the solutions of the u equation are approximated by the solutions of $u''(x) - u(x) = 0$; i.e. $u_1(x) = A_0e^{x}$ and $u_2(x) = B_0e^{-x}$. Use this in conjunction with the change of variable in (6.5.11) to find the leading terms of the solutions $y_1(x)$ and $y_2(x)$ displayed in Example 6.5.1.

6.6 WKB Methods

When $g(x)$ in (6.5.13) is one-signed throughout the interval of interest a change of independent variable followed by a change of dependent variable brings (6.5.1) to one of the forms

$$y''(x) - [1 + f(x)]y(x) = 0 \tag{6.6.1}$$

or

$$y''(x) + [1 + f(x)]y(x) = 0 \tag{6.6.2}$$

where $f(x) \to 0$, $x \to \infty$. The first form (6.6.1) is reached when $g(x)$ in (6.5.13) is negative and the second (6.6.2) when $g(x)$ is positive. For the moment, we suppose that the requisite changes of variable have been made (the connection with (6.5.13) is established later).

Consider, for example, (6.6.1) which is analogous to the simple equation

$$y'' - y = 0\,.$$

This equation has the solutions

$$y = \exp(\pm x) = \exp\left[\pm \int^x 1\, dt\right].$$

Using this analogy, we see that a solution of (6.6.1) has the form

$$y = \exp\left[\int^x (1 + \eta(t))dt\right] \tag{6.6.3}$$

so that

$$y' = [1 + \eta(x)]y$$

and

$$y'' = [(1 + \eta)^2 + \eta']y\,.$$

Equation (6.6.1) is satisfied if

$$\eta' + (1 + \eta)^2 = 1 + f(x)\,. \tag{6.6.4}$$

If we assume that η can be expressed as the series

$$\eta = \eta_1 + \eta_2 + \eta_3 + \cdots$$

then (6.6.4) becomes

$$1 + 2\eta_1 + (2\eta_2 + \eta_1^2) + (2\eta_3 + 2\eta_1\eta_2) + \cdots = 1 + f - \eta_1' - \eta_2' - \eta_3' - \cdots$$

and these equations are soluble for the η_i if they are arranged in the form

$$\begin{aligned} 2\eta_1 &= f \\ 2\eta_2 &= -\eta_1^2 - \eta_1' \\ 2\eta_3 &= -2\eta_1\eta_2 - \eta_2' . \end{aligned}$$

From these, we obtain in sequence

$$\begin{aligned} \eta_1 &= \frac{1}{2} f \\ \eta_2 &= -\frac{1}{4} f' - \frac{1}{8} f^2 \\ \eta_3 &= \frac{1}{4} f f' + \frac{1}{8} f'' + \frac{1}{16} f^3 . \end{aligned}$$

If the derivatives of f approach zero, $x \to \infty$, a few terms of this series may yield a satisfactory asymptotic approximation. Since each of the terms η_i is to be integrated in the expression (6.6.3), it is not possible to effect the integration explicitly until the form of $f(x)$ is given.

Clearly, (6.6.1) has a second solution of the form

$$y = \exp\left[\int^x (-1 + \eta(t))dt\right]$$

but it is simpler to obtain the form of this solution by the method of Example 6.6.1 below.

Example 6.6.1 A solution of the equation

$$y'' = a^2[1 + f(x)]y$$

is of the form

$$y = \exp\left[\int^x \big(a + \eta(t)\big)dt\right]$$

where

$$\eta = \eta_1 + \eta_2 + \eta_3 + \cdots .$$

Verify that

$$\begin{aligned} \eta_1 &= \frac{af}{2} \qquad \eta_2 = -\frac{f'}{4} - \frac{af^2}{8} \\ \eta_3 &= \frac{ff'}{4} + \frac{f''}{8a} + \frac{af^3}{16} . \end{aligned}$$

The second solution of equations (6.6.1) can now be obtained on replacing a by -1. Similarly, the two (complex) solutions of (6.6.2) are obtained on replacing a by i and $-i$.

Example 6.6.2 The equation

$$y'' = a^2 \left[1 + \frac{b}{x^2}\right] y$$

is a special case of (6.6.1) or (6.6.2), but is of some generality. It can be shown that the Stokes-type solution of this equation is

$$\begin{aligned} y \quad = \quad & A_0 e^{ax}\left[1 - \frac{ab}{2x} - \frac{b(2-a^2b)}{8x^2} - \frac{b(2-a^2b)(6-a^2b)}{48ax^3} - \cdots\right] \\ & + B_0 e^{-ax}\left[1 + \frac{ab}{2x} - \frac{b(2-a^2b)}{8x^2} + \frac{b(2-a^2b)(6-a^2b)}{48ax^3} - \cdots\right] . \end{aligned}$$

Example 6.6.3 Verify from Example 6.6.1 that

$$y'' = \left(1 + \frac{1}{x^2}\right) y$$

has the solution

$$y = \exp\left[x - \frac{1}{2x} - \frac{1}{4x^2} - \cdots\right]$$

and verify from Example 2 that the solution can be written as

$$y = e^x \left[1 - \frac{1}{2x} - \frac{1}{8x^2} - \cdots\right] .$$

Verify that these solutions are consistent up to the terms in $1/x^2$ (Problem 6.6.4).

Example 6.6.4 Verify that

$$y_1(x) = \exp\left(x - \frac{1}{2x}\right)$$

satisfies the equation

$$y_1'' = \left(1 + \frac{1}{x^2}\right) y_1 + r(x)$$

where

$$r(x) = -\left(\frac{1}{x^3} - \frac{1}{4x^4}\right) y_1 .$$

The device used for the approximate solution of (6.6.1) and (6.6.2) can be used with advantage for the more general equation (6.5.13). For a reason that appears presently, we replace (6.5.13) by

$$y'' = \frac{g(x)}{h^2} y , \quad g(x) > 0 , \quad a \le x \le b \tag{6.6.5}$$

where we assume that $g(x)$ is positive throughout the interval where the solution is to be determined and h is a parameter that we may choose freely. In particular, putting $h = \pm 1$, we obtain the equation

$$y'' = g(x)y$$

in which the multiplier of y on the right is a positive function; again, putting $h = \pm i$, we obtain the equation

$$y'' = -g(x)y$$

in which the multiplier of y on the right is a negative function. For the moment, it is convenient to regard h as a small positive parameter.

We seek a solution of (6.1.5) of the form

$$y = \exp\left[\frac{1}{h}\int^{x}(\mu_0 + h\mu_1 + h^2\mu_2^2 + \cdots)dt\right] \tag{6.6.6}$$

where

$$\mu_i = \mu_i(t) \quad , \quad i = 0, 1, 2, \ldots$$

is a differentiable function of t. Proceeding as before, we find the equation

$$\frac{1}{h^2}[\mu_0 + h\mu_1 + h^2\mu_2 + \cdots]^2 + \frac{1}{h}[\mu_0' + h\mu_1' + h^2\mu_2' + \cdots] = \frac{g(x)}{h^2}$$

where the primes denote differentiation with respect to x. Equating the coefficients of powers of h on both sides of this equation, we arrive at the sequence of equations

$$\begin{aligned} \mu_0^2 &= g \\ 2\mu_0\mu_1 + \mu_0' &= 0 \\ 2\mu_0\mu_2 + \mu_1^2 + \mu_1' &= 0 \\ \cdots & \end{aligned} \tag{6.6.7}$$

The first three equations yield

$$\begin{aligned} \mu_0 &= \pm g^{1/2} \qquad \mu_1 = -\frac{1}{2}\frac{\mu_0'}{\mu_0} \\ \mu_2 &= \frac{1}{4\mu_0}\left[\frac{\mu_0''}{\mu_0} - \frac{3}{2}\frac{(\mu_0')^2}{\mu_0^2}\right] \; . \end{aligned} \tag{6.6.8}$$

If we denote by y_1 the approximation obtained by retaining the terms in μ_0 and μ_1 in (6.6.6), then

$$y_1(x) \quad = \exp\left[\tfrac{1}{h}\int^{x}\left\{\mu_0(t) - \tfrac{h}{2}\tfrac{\mu_0'(t)}{\mu_0(t)}\right\}dt\right]$$

$$= |\mu_0|^{-1/2} \exp\left[\tfrac{1}{h}\int^x \mu_0(t)dt\right] \tag{6.6.9}$$

where $\mu_0 = \pm g^{1/2}$. The function $y_1(x)$ is called the WKB approximation. We surmise that the solution $y_1(x)$ is related to the true solution $y(x)$ by the relation

$$y(x) = y_1(x)\left[1 + 0\Big(\exp\Big[h\int^x \eta_2(t)dt\Big]\Big)\right].$$

This surmise cannot be confirmed by the present analysis since the convergence of the series μ_i cannot be demonstrated. We can however establish that $y_1(x)$ satisfies a differential equation whose coefficients are close to the original equation. For

$$\begin{aligned} y_1(x) &= \exp\left[\frac{1}{h}\int^x \left\{\mu_0 - \frac{h}{2}\frac{\mu_0'}{\mu_0}\right\} dt\right] \\ \frac{d^2y_1}{dx^2} &= \left[\frac{\mu_0^2}{h^2} - 2\mu_2\right] y_1 \\ &= \left[\frac{g(x)}{h^2} - 2\mu_2\right] y_1\,. \end{aligned} \tag{6.6.10}$$

For h small, we observe that the second term in brackets above is independent of h and is relatively small, $h \to 0$. For $h = 1$ we require

$$\frac{2\mu_2}{\mu_0^2} = \frac{2\mu_2}{g} \ll 1\,.$$

It is important to realize that (6.6.9) embraces two solutions since

$$\mu_0 = \pm g^{1/2}$$

and we know from the general theory that a second-order differential equation possesses two independent solutions.

Example 6.6.5 When the function g in (6.6.5) given by

$$g(x) = 1 - \frac{1}{x^2+1}$$

we see from (6.6.8) that

$$\begin{aligned} \mu_0 &= \pm\tfrac{x}{(1+x^2)^{1/2}} \\ \mu_2 &= \mp\tfrac{3}{4(1+x^2)^2}\left[1 \pm \tfrac{1}{2x^2}\right] . \end{aligned}$$

For $x \geq 10$ then

$$|\mu_2| < 7 \cdot 5 \times 10^{-5} .$$

Example 6.6.6 When

$$\begin{aligned} g(x) &= e^{2x} \qquad\qquad \mu_0 = \pm e^x \\ \frac{|\mu_2|}{g(x)} &= \frac{1}{8}e^{-2x} \\ &< 6 \times 10^{-6} \text{ for } x \geq 5 \end{aligned}$$

Example 6.6.7 Taking $h = \pm i$ then the WKB approximation of the equation

$$y'' = -g(x)y \quad , \quad g(x) > 0$$

is

$$y_1(x) = |\mu_0|^{-1/2} \exp\left[\mp i \int^x \mu_0(t)dt\right] .$$

Example 6.6.8 It is sometimes desirable to express the WKB approximation explicitly in terms of g. Equation (6.6.9) can be expressed as

$$y_1(x) = |g(x)|^{-1/4} \exp\left[\frac{1}{h}\int^x g^{1/2}(t)dt\right] \tag{6.6.11}$$

and

$$\mu_2 = \frac{1}{8}\left[\frac{g''}{g} - \frac{5}{4}\frac{(g')^2}{g^2}\right] . \tag{6.6.12}$$

The equation (6.6.10) becomes

$$\frac{d^2y_1}{dx^2} = \left[\frac{g(x)}{h^2} - \frac{1}{4}\left(\frac{g''}{g} - \frac{5}{4}\frac{(g')^2}{g^2}\right)\right] y_1 \tag{6.6.13}$$

but the earlier notation seems preferable.

To begin the analysis implied after (6.6.9) we consider

$$y'' \pm [1 + f(x)]y = 0$$

and subsequently extend the analysis to the more general equations

$$y'' \pm g(x)y = 0 \quad , \quad g(x) > 0 .$$

It is convenient to consider positive and negative signs separately, so in this direction we state two separate but related lemmas.

Lemma 6.6.9 *If $y(x)$ satisfies the differential equation*

$$y'' + [1 + f(x)]y = 0 \quad , \quad a \le x \le b \tag{6.6.14}$$

and if we can find a $w(x)$ that satisfies the equation

$$w'' + [1 + f(x)]w = r(x) \quad , \quad a \le x \le b \tag{6.6.15}$$

where

$$\int_a^b |f(t)|dt < \infty \quad , \quad \int_a^b |r(t)|dt < \infty \, , \tag{6.6.16}$$

then there exists a solution $y(x)$ of (6.6.14) such that

$$|y(x) - w(x)| \le \exp\left[\int_x^b |f(t)|dt\right] \int_x^b |r(t)|dt \, . \tag{6.6.17}$$

Proof: On subtracting equation (6.6.15) from (6.6.14) and writing

$$\epsilon = y - w \tag{6.6.18}$$

we find that ϵ satisfies the differential equation

$$\epsilon'' + [1 + f(x)]\epsilon = -r(x) \, . \tag{6.6.19}$$

It can be verified that $\epsilon(x)$ defined by (6.6.19) satisfies the integral equation

$$\epsilon(x) = \int_x^b \sin(x - t)[f(t)\epsilon(t) + r(t)]dt \ , \quad a \le x < b \tag{6.6.20}$$

and it is much easier to obtain a bound on $\epsilon(x)$ from the integral equation. It can also be verified that

$$\epsilon(x) \to 0 \quad , \quad \epsilon'(x) \to 0 \ , \quad x \to b$$

and from these properties it is clear that $\epsilon(x)$ defined by (6.6.20) implies $y(x) \to w(x)$, $y'(x) \to w'(x)$, $x \to b$. Since

$$|\sin(x - t)| \le 1$$

it follows that

$$|\epsilon(x)| \le \int_a^b [\,|f(t)|\ |\epsilon(t)| + |r(t)|\,]\,dt \, .$$

Let us define $E(x)$ by

$$E(x) = \int_x^b [\,|f(t)|E(t) + |r(t)|\,]\,dt\,.$$

Differentiating $E(x)$ with respect to x, we find that $E(x)$ satisfies the first-order equation

$$E'(x) = -|f(x)|E(x) - r(x) \quad , \quad E(b) = 0\,.$$

Writing

$$F(x) = \int_x^b |f(t)|dt$$

the solution of this differential equation is

$$E(x) = \int_x^b \exp[F(x) - F(t)]|r(t)|dt$$

and it is now clear that $E(x)$ is positive in $a \le x \le b$ and that

$$|\epsilon(x)| \le E(x)\,.$$

The expression for the error can be simplified slightly; since $F(x)$ is positive and decreasing in $a \le x \le b$, we have

$$F(x) - F(t) \le F(x)\,.$$

Hence finally

$$|\epsilon(x)| \le \exp(F(x)) \int_x^b |r(t)|dt$$

where

$$F(x) = \int_x^b |f(t)|dt\,. \tag{6.6.21}$$

This is the inequality (6.6.17).

The existence of a bounded solution $\epsilon(x)$ implies the existence of a bounded solution

$$y(x) = w(x) + \epsilon(x)$$

of (6.6.14). Notice that (6.6.15) has two independent solutions and correspondingly there are two solutions of (6.6.14).

The corresponding lemma for the equation

$$y'' = [1 + f(x)]y \quad , \quad a \le x \le b \tag{6.6.22}$$

may be formulated thus:

Lemma 6.6.10 *If $w_1(x)$ and $w_2(x)$ satisfy the differential equations*

$$\begin{aligned} \frac{d^2 w_1}{dx^2} &= [1+f(x)]w_1 + r_2 e^x \quad , \quad a \le x \le b \\ \frac{d^2 w_2}{dx^2} &= [1+f(x)]w_2 + r_2 e^{-x} \quad , \quad a \le x \le b \end{aligned} \tag{6.6.23}$$

and if

$$\int_a^b |f(t)|dt < \infty \quad , \quad \int_a^b |r_i(t)|dt < \infty \quad , \quad i = 1, 2 \, .$$

Then, in the notation,

$$\begin{aligned} w_1 &= e^x v_1 & w_2 &= e^{-x} v_2 \\ y_1 &= e^x z_1 & y_2 &= e^{-x} z_2 \end{aligned}$$

There exist solutions y_1 and y_2 of (6.6.22) such that

$$|z_1(x) - v_1(x)| \le \frac{1}{2} \exp\left[\int_a^x |f(t)|dt\right]\left(\int_a^x |r_1(t)|dt\right)$$

and

$$|z_2(x) - v_2(x)| \le \frac{1}{2} \exp\left[\int_x^b |f(t)|dt\right]\left(\int_x^b |r_2(t)|dt\right) .$$

The proof of this lemma is similar to the proof of the preceding lemma and may be omitted; the principal steps of the proof are suggested in the problems that follow.

Problems

6.6.1 If w_1 satisfies the first of (6.6.23) and if

$$\epsilon_1 = y_1 - w_1$$

show that ϵ_1 satisfies the differential equation

$$\epsilon_1'' = [1+f(x)]\epsilon_1 - r_1 e^x$$

and the integral equation

$$\epsilon_1 = \int_b^x \sinh(x-t)\left[f(t)\epsilon_1(t) - r_1(t)e^t\right] dt \, .$$

Writing $\epsilon_1 = \eta_1 e^x$, show that η_1 satisfies the integral equation

$$\eta_1(x) = \int_a^x e^{-(x-t)} \sinh(x-t) \, [f(t)\eta_1 - r_1(t)] \, dt \, .$$

Note that

$$|e^{-(x-t)} \sinh(x-t)| \leq \frac{1}{2} \, , \quad t \leq x$$

and that the lower limit a may be taken to be $-\infty$.

6.6.2 If w_2 satisfies the second of (6.6.23) and if

$$\epsilon_2 = y_2 - w_2$$

show that ϵ_2 satisfies the equations

$$\begin{aligned} \epsilon_2'' &= [1 + f(x)]\epsilon_2 - r_2 e^{-x} \\ \epsilon_2 &= \int_x^b \sinh(t-x) \left[f(t)\epsilon_2(t) - r_2(t)e^{-t} \right] dt \, . \end{aligned}$$

Writing $\epsilon_2 = \eta_2 e^{-x}$, show that

$$\eta_2(x) = \int_x^b e^{-(t-x)} \sinh(t-x) \, [f(t)\eta_2 - r_2(t)] \, dt \, .$$

The limit b may be taken to be $+\infty$.

6.6.3 If

$$y'' = \left[1 + \frac{1}{x^m} \right] y$$

where m is an integer ≥ 2, show that the coefficients of the asymptotic series

$$y = e^x \sum_{k=0}^{\infty} \frac{A_k}{x^k}$$

are determined by the equations

$$\begin{aligned} 0 &= \frac{-2(m-1)A_{m-1} - A_0}{x^m} - 2\sum_{k=0}^{\infty} \frac{(m+k)A_{m+k}}{x^{m+k+1}} \\ &\quad + \sum_{k=0}^{\infty} \frac{(m+k-1)(m+k)A_{m+k+1}}{x^{m+k+1}} - \sum_{k=0}^{\infty} \frac{A_{m+k}}{x^{m+k}} \end{aligned}$$

and that the non-zero coefficients are determined by

$$\begin{aligned}
A_{m-1} &= -\frac{1}{2(m-1)}\,A_0 \\
A_{m+k} &= \frac{1}{2}\,(m+k-1)A_{m+k-1}\ , \quad k=0,1,2,\ldots,m-3 \\
A_{2m+k-2} &= \frac{1}{2}\,(2m+k-3)A_{2m+k-3} \\
&\quad -\frac{1}{2(2m+k-3)(2m+k-2)}\,A_{m+k-1}\ , \quad k=0,1,2,\ldots
\end{aligned}$$

Further, apply the preceding to the differential equation with $m = 4$ to show that the relative error of the approximation

$$y_1(x) = A_0 e^x \left[1 - \frac{1}{6x^3}\right]$$

is less than

$$\frac{1}{2}\,A_0 \left\{\frac{1}{24x^4} + \frac{1}{6x^6}\right\} \exp\left(\frac{1}{6x^3}\right)\ .$$

6.6.4 Referring to Example 3 verify the steps

$$\begin{aligned}
y(x) &= \exp\left[x - \frac{1}{2x} - \frac{1}{4x^2} - \cdots\right] \\
&= e^x\left[\left(1 - \frac{1}{2x} + \frac{1}{2(4x)^2}\ \cdots\right)\left(1 - \frac{1}{4x^2} + \frac{1}{2(4x^2)^2}\ \cdots\right)\cdots\right] \\
&= e^x\left[1 - \frac{1}{2x} - \frac{1}{8x^2} - \cdots\right]
\end{aligned}$$

so that the consistency of the last line of that example is established.

6.7 Liouville's Transformation

It remains to bring the equations

$$\begin{aligned}
y'' + g(x)y &= 0 \qquad\qquad g(x) > 0\ , \quad a \le x \le b \\
y'' - g(x)y &= 0 \qquad\qquad g(x) > 0\ , \quad a \le x \le b
\end{aligned} \tag{6.7.1}$$

within the scope of the error-bound analysis of the preceding section; we consider first the first equation of (6.7.1).

It is convenient to retain the earlier notation so that μ_0 denotes either the positive or negative square root of $g(x)$ and to write

$$\mu = +g^{1/2}(x)\,. \tag{6.7.2}$$

The change of independent variable

$$\xi = \int_a^x \mu(t)dt \tag{6.7.3}$$

defines a one-to-one map of the interval $a < x < b$ onto the interval $0 < \xi < c$ where

$$c = \int_a^b \mu(t)dt\,. \tag{6.7.4}$$

From (6.7.3)

$$\frac{d\xi}{dx} = \mu(x)$$

and

$$\begin{aligned} \frac{dy}{dx} &= \mu(x)\,\frac{dy}{d\xi} \\ \frac{d^2y}{dx^2} &= \mu^2\,\frac{d^2y}{d\xi^2} + \frac{d\mu}{dx}\,\frac{dy}{d\xi}\,. \end{aligned}$$

The first of (6.7.1) becomes, on division by μ^2,

$$\frac{d^2y}{d\xi^2} + \frac{1}{\mu^2}\,\frac{d\mu}{dx}\,\frac{dy}{d\xi} + y = 0\,. \tag{6.7.5}$$

The middle term in the above equation can be removed by the change of dependent variable (6.5.11). It is convenient to write temporarily

$$\begin{aligned} \mu(x) &= \lambda(\xi) \\ \frac{d\mu}{dx} &= \lambda\,\frac{d\lambda}{d\xi}\ ,\quad \frac{1}{\mu^2}\,\frac{d\mu}{dx} = \frac{1}{\lambda}\,\frac{d\lambda}{d\xi}\,. \end{aligned} \tag{6.7.6}$$

We can now make the change of dependent variable

$$\begin{aligned} y &= u\exp\left[-\frac{1}{2}\int^\xi \frac{1}{\lambda(t)}\,\frac{d\lambda(t)}{dt}\,dt\right] \\ &= \lambda^{-1/2}(\xi)u \end{aligned}$$

and, from (6.5.12), the equation in u is

$$\frac{d^2u}{d\xi^2}+\left[1-\frac{1}{2}\left\{\frac{1}{\lambda}\frac{d^2\lambda}{d\xi^2}-\frac{1}{2\lambda^2}\left(\frac{d\lambda}{d\xi}\right)^2\right\}\right]u=0 \qquad (6.7.7)$$

which is of the form (6.6.2). Hence the error-bound (6.6.17) may be used for the approximate solutions of (6.7.7) and hence of the first of (6.7.1) since y is related to u by known transformations. Writing

$$\mu(t)=\lambda(s)$$

and

$$\frac{ds}{dt}=\mu(t)$$

the expression $F(\xi)$ in (6.6.21) can be expressed in terms of t and x. We find

$$\begin{aligned} F(\xi) &= \frac{1}{2}\int_x^b\frac{1}{\mu}\left[\frac{1}{\mu}\frac{d^2\mu(t)}{dt^2}-\frac{3}{2\mu^2}\left(\frac{d\mu(t)}{dt}\right)^2\right]dt \\ &= 2\int_x^b\frac{|\mu_2(t)|}{\mu(t)}\,dt \end{aligned} \qquad (6.7.8)$$

in the notation of (6.6.8) if μ_2 as there defined is replaced by μ.

Example 6.7.1 Let us rewrite (6.7.7) in the form

$$\frac{d^2u}{d\xi^2}+[1+h(\xi)]u=0 \qquad (6.7.9)$$

where

$$h(\xi)=-\frac{1}{2}\left[\frac{1}{\lambda}\frac{d^2\lambda}{d\xi^2}-\frac{1}{2\lambda^2}\left(\frac{d\lambda}{d\xi}\right)^2\right].$$

The function

$$u_1=\exp(\pm i\xi)$$

satisfies the equation

$$\frac{d^2u_1}{d\xi^2}+[1+h(\xi)]u_1=\rho(\xi) \qquad (6.7.10)$$

where

$$\rho(\xi)=h(\xi)u_1$$

in the notation of (6.6.8) if μ_2 as there defined is replaced by μ. Also

$$\int_\xi^c|\rho(s)|ds=2\int_x^b\frac{|\mu_2(t)|}{\mu(t)}\,dt$$

where the left-hand side is the integral of Problem 6.6.1.

Further Reading

N. Bleistein and R.A. Handelsman. *Asymptotic Expansions of Integrals*. Dover Publications, Inc., New York, 1986.

N.G. deBruijn. *Asymptotic Methods in Analysis*. Dover Publications, Inc., New York, 1981.

A. Erdélyi. *Asymptotic Expansions*. Dover Publications, Inc., New York, 1956.

F.W.J. Olver. *Asymptotics and Special Functions*, Academic Press, New York (1974).

Chapter 7

Numerical Solution of Ordinary Differential Equations

Introduction

We assume in this chapter that the reader is familiar with the elementary theory of ordinary differential equations. Hence, although we state a few of the theorems governing existence and uniqueness of solutions we neither prove them nor claim that the list is exhaustive. Many excellent textbooks exist on this subject, some we list at the end of the chapter. The reader should note as she proceeds through this chapter that the flavor when numerically solving initial values problems versus boundary value problems is very different.

When solving initial value problems, we may, without loss of generality, restrict ourselves to first order nonlinear systems of the form

$$y' = f(x, y) \ , \ \ y(x_0) = y_0 \ , \tag{7.0.1}$$

where y, y_0, x, and x_0 are vectors if appropriate. This is possible since nth order equations can always be transformed into a system of such equations. Many methods are known for solving (7.0.1). We discuss and illustrate by example only the simpler among these, such as Euler's method, Heun's method, Runge-Kutta methods, and predictor-corrector methods. These are still typically applied when numerically solving partial differential equations that evolve in time. We also carry out an error analysis, discussing the computational difficulties that may arise due to instability and roundoff error.

Finally, we consider the numerical solution of stiff equations. These commonly occur for systems that model a physical process with widely varying time constants. For example, a chemical system may be dominated for a short time by one very fast reaction followed by a second reaction that is extremely slow. Care must be taken when selecting a method for the system otherwise the fast reaction dictates an extremely small step size to maintain numerical stability. Quite likely, the small step size in incompatible with monitoring the behavior of the slower second phase.

For the case of the numerical solution of boundary value problems, we restrict ourselves to the second order equation

$$y'' + f(x, y, y') = 0 \ , \quad y(0) = y(1) = 0 \, .$$

Here we again describe a selected group of the well-known simple methods that lead to quickly obtainable approximations. Included among these are Numerov's method, the shooting method combined with initial value methods, and methods that lead to the solution of a system of algebraic equations. This latter class includes finite difference methods and techniques that use "elements" such as cubic splines. We also discuss the numerical solution of eigenvalue problems and the Bubnov-Galerkin method.

The topic of this chapter presents unusual difficulties to students. It makes extensive use of numerical integration and, in addition, calls on concepts from two fields: differential equations and difference equations. There is a vast literature of differential equations and an almost equally vast literature of numerical solutions of these equations. If this chapter serves to orient the reader in his subsequent study of the literature, it has served its purpose.

7.1 Illustrative Examples of Ordinary Differential Equations

In this section we discuss examples in order to elucidate essential points of the theory of differential equations, and to illustrate the kind of preliminary analysis that often precedes numerical solution. Throughout this chapter a prime denotes differentiation with respect to x.

The simplest first order linear differential equation,

$$y' = f(x) \, , \tag{7.1.1}$$

immediately admits a formal integral solution

$$y(x) = y_0 + \int_{x_0}^{x} f(t)dt \tag{7.1.2}$$

where $y(x_0) = y_0$. Once the point $y(x_0) = y_0$ is prescribed, the solution $y(x)$ of (7.1.1) is unique. Similarly, n auxiliary conditions must be prescribed for an equation of order n (e.g. the equation $y^{(n)}(x) = f(x)$) to have a unique solution. Usually the integral in (7.1.2) cannot be explicitly expressed. We must therefore resort to a method of numerical integration as described in Chapter 5 in order to approximate $y(x)$.

The general form of a first order linear differential equation is

$$y' = a(x)y + f(x)\,. \tag{7.1.3}$$

The circumstances gauranteeing existence and uniqueness are specified in the following theorem.

Theorem 7.1.1 *If a and f are continuous in the interval $\alpha < x\beta$, and then there exists a unique differentiable solution $y = y(x)$ of (7.1.3) defined on (α, β) and satisfying $y(x_0) = y_0$.*

The solution of this equation is

$$y(x) = \exp\{A(x)\}\left[y_0 + \int_{x_0}^{x} \exp\{-A(t)\}f(t)dt\right] \tag{7.1.4}$$

where

$$A(x) = \int_{x_0}^{x} a(t)dt\ ,\ \ y_0 = y(x_0)\,. \tag{7.1.5}$$

From a standpoint of analysis, the solution is complete. However, it may be inconvenient for purposes of numerical evaluation. For example, it may not be possible to explicitly express either $A(x)$, or the integral

$$\int_{x_0}^{x} \exp\{-A(t)\}f(t)dt\,. \tag{7.1.6}$$

In addition, the numerical evaluation of these integrals may be difficult if either $a(t)$ or $f(t)$ has a singularity on the interval $[x_0, x]$ (see Problem 7.1.6 and Chapter 5) since the standard quadrature rules of Chapter 5 require more than just continuity of the integrand. If neither $A(t)$ nor the integral (7.1.6) can be expressed explicitly, it may be much simpler to use a direct method from this chapter as an approximation technique.

Example 7.1.2 Consider the differential equation

$$xy' + 2y = 4x^2 \ , \ \ y(1) = 1 + \varepsilon \, . \tag{7.1.7}$$

Rewriting the equation to match the form of (7.1.3) we see that $a(x) = -2/x$ from which simple integrations yield $A(x) = -2 \log x$ and

$$y(x) = x^2 + \frac{\varepsilon}{x^2} \, .$$

If $\varepsilon = 0$ then this solution is valid for all x otherwise we must take care to exclude $x = 0$ from consideration. However, even in the former case the numerical student should expect difficulties due to round off error.

Next we turn our attention to nonlinear first order equations of the form

$$y' = f(x, y) \ , \ \ y(x_0) = y_0 \, . \tag{7.1.8}$$

These problems differ significantly from the linear case. Perhaps most important for our discussion is that we can not write down a general analytic solution such as (7.1.4). Hence approximation techniques take an even a more important role.

Theorem 7.1.3 *If $f(x, y)$ and $\partial f/\partial y$ are continuous functions of x and y in a rectangle R given by $\alpha < x < \beta$ and $\gamma < y < \delta$ where $(x_0, y_0) \in R$ then a unique solution of (7.1.8) exists in an interval $(x_0 - h, x_0 + h) \subseteq (\alpha, \beta)$.*

Example 7.1.4

$$y' = x^2 + y^2 \tag{7.1.9}$$

By Theorem 7.1.3 we know that a unique solution of the differential equation exists. Indeed, the solution can be expressed explicitly (but inconveniently) in terms of Bessel functions. The isoclines

$$y' = \text{const} \tag{7.1.10}$$

aid in graphing solution curves. Here they are the circles

$$\{(x, y) : x^2 + y^2 = c^2\} \tag{7.1.11}$$

that are graphed in Figure 7.1.1 for $c_1 < 1 < c_2$. The solution curves passing through the x-axis at x_0 have the slopes

$$y'(x_0) = x_0^2 \, . \tag{7.1.12}$$

Numerical difficulties may be expected in the region

$$\{(x, y) : x^2 + y^2 > 1\} \tag{7.1.13}$$

where the slopes of the solution curves of (7.1.9) are very steep.

Figure 7.1.1: Isoclines of $x^2 + y^2 = c^2$.

Example 7.1.5 For an example not addressed by Theorem 7.1.3 consider the nonlinear differential equation

$$(y')^2 = \frac{1-y^2}{1-x^2}\, a(x)\ , \quad x_0 \leq x \leq x_1 \tag{7.1.14}$$

where a is a continuous function of x on the interval $[x_0, x_1]$. Except for particular functions a, the solutions of (7.1.14) cannot be explicitly expressed.

A nonlinear differential equation that is linear in the highest derivative is called quasi-linear. If we take $a(x) = 1$ and then take the square root of each side of (7.1.14) we obtain two quasi-linear equations, namely

$$y' = \left(\frac{1-y^2}{1-x^2}\right)^{1/2} \quad \text{and} \quad y' = -\left(\frac{1-y^2}{1-x^2}\right)^{1/2} . \tag{7.1.15}$$

These equations are defined only in the regions

$$\begin{aligned} R_1 &= \{(x,y) : x^2 \leq 1\ ,\ y^2 \leq 1\} \\ R_2 &= \{(x,y) : x^2 \geq 1\ ,\ y^2 \geq 1\}\,. \end{aligned}$$

In R_1, the solutions of (7.1.15) are a one-parameter family of hyperbolas. For example, with reference to Figure 7.1.2, the solid curves are solutions to the first equation in (7.1.15) while the dotted curves are solutions to the second equation in (7.1.15) (solutions given by $x^2 - 2cxy + y^2 = 1 - c^2$ where $y(0) = 1 - c^2$). The solution curves are tangent to the boundary of R_1 at A, B, C, and D, so that starting from D and proceeding to C along the curve, y' increases from 0 to ∞. This property makes it difficult to solve (7.1.15) numerically.

Figure 7.1.2:

We next turn our attention to the general second order linear differential equation

$$y'' + a(x)y' + b(x)y = c(x) \ , \ \ x_0 < x < x_1 \tag{7.1.16}$$

$$y(x_0) = y_0 \quad , \quad y'(x_0) = y_0' \tag{7.1.17}$$

where a, b and c are continuous functions defined on (x_0, x_1) .

Theorem 7.1.6 *If $a(x)$, $b(x)$, and $c(x)$ are continuous functions of x on (x_0, x_1) then the solution of (7.1.16) exists on (x_0, x_1) and is unique.*

Despite its apparent simplicity, it is seldom possible to express the general solution of (7.1.16) in a form such as (7.1.4), involving only a finite number of integrations. If $c(x) = 0$ on $[x_0, x_1]$, then (7.1.16) is said to be homogeneous. Otherwise it is said to be nonhomogeneous. Once a nontrivial solution ($y(x) \neq 0$ on $[x_0, x_1]$) of the homogeneous equation is known, we can reduce (7.1.16) to a first order linear equation, i.e., we can explicitly express a solution in closed form (see Problem 7.1.3). The homogeneous form of (7.1.16) is

$$y'' + a(x)y' + b(x)y = 0\,.$$

The presence of $a(x)y'$ in (7.1.16) may be inconvenient in numerical work and it may be advantageous to transform (7.1.16) into another second order linear equation where there is no first derivative term present. This may be achieved in the homogeneous case via

$$y = v\exp\left[-\frac{1}{2}\,A\right]\ ,\ A(x) = \int_{x_0}^{x} a(t)dt \tag{7.1.18}$$

so that

$$v'' + g(x)v = 0\ ;\ g(x) = b(x) - \frac{1}{2}\,a'(x) - \frac{1}{4}\,a^2(x)\,. \tag{7.1.19}$$

If (7.1.16) is to be solved subject to the initial condition

$$y(x_0) = y_0\ ,\ y'(x_0) = y_0'$$

then the corresponding initial conditions for v become

$$v(x_0) = y_0\ ,\ v'(x_0) = y_0' + \frac{1}{2}\,a(x_0)y_0\,.$$

The solution of the problem (7.1.19) can be expressed in series form

$$v(x) = \sum_{n=0}^{\infty} v_n(x) \tag{7.1.20}$$

where

$$\begin{aligned} v_0(x) &= v(x_0) + v'(x_0)(x - x_0) \\ v_n(x) &= \int_{x_0}^{x}(t - x)g(t)v_{n-1}(t)dt\ ,\ \ n = 1, 2, 3\ldots \end{aligned} \tag{7.1.21}$$

Notice the linear dependence of the solution on the initial values. The procedure (7.1.20)–(7.1.21) may be a worthwhile procedure in applications if $|g(t)|$ and the length of interval $[x_0, x_1]$ are not too large. For if

$$|g(t)| \le M\ ,\ \ |v_0(t)| \le \mu \tag{7.1.22}$$

on the interval $[x_0, x_1]$, then it may be shown (Problem 7.1.2) that

$$|v_n(x)| \le \frac{\mu M^n |x - x_0|^{2n}}{(2n)\,!} \tag{7.1.23}$$

Example 7.1.7 The equation

$$y'' + c(x)y = 0 \;, \quad x \le x \le x_1 \,, \tag{7.1.24}$$

where c is a large, positive and slowly varying function of x, has solutions that oscillate rapidly. It is difficult to obtain an accurate solution of this equation by standard numerical methods. This difficulty can be eliminated by the Madelong transformation

$$y = u(x)\cos[v(x)]\,. \tag{7.1.25}$$

In this transformation, y is expressed in terms of its modulus $u(x)$ and phase $v(x)$. Differentiating (7.1.25), we get

$$\begin{aligned} y' &= u'\cos(v) - uv'\sin(v) \\ y'' &= [u'' - u(v')^2]\cos v - [uv'' + 2u'v']\sin v \end{aligned}$$

If we substitute these expressions into the original expression and require that the coefficients of $\cos v$ and $\sin v$ vanish separately, we obtain the two equations

$$\begin{aligned} uv'' + 2u'v' &= 0 \\ u'' + [c - (v')^2]v &= 0 \end{aligned} \tag{7.1.26}$$

The first equation in (7.1.26) can be integrated at once, to yield

$$v'(x) = \frac{k}{u^2}$$

where k is a constant, so that

$$v(x) = k\int_{x_0}^{x} \frac{dt}{u^2(t)}\,. \tag{7.1.27}$$

The second equation in (7.1.26) then becomes

$$u'' + \left[c - \frac{k^2}{u^4}\right] u = 0\,. \tag{7.1.28}$$

If (7.1.24) is to be solved subject to the initial conditions

$$y(x_0) = y_0 \;, \quad y'(x_0) = y_0'$$

then the corresponding initial conditions for (7.1.28) become

$$u(x_0) = y_0 \;, \quad u'(x_0) = y_0'\,.$$

We may now choose the constant k^2 in (7.1.28) to make this equation simpler to solve numerically. One such choice is

$$k^2 = c(x_0)y_0^4\,.$$

The WKB method described in Chapter 6 may also be used to obtain an approximate solution of (7.1.24).

Example 7.1.8 An n'th order equation in one dependent variable can always be rewritten as a system of n first order equations. For example, in the Van der Pol equation

$$y'' + \mu(y^2 - 1)y' + y = 0 \tag{7.1.29}$$

where μ is a constant, we rename y as y_1 and y' as y_2. Then (7.1.29) becomes

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= \mu(y_1^2 - 1)y_2 - y_1 \end{aligned} \tag{7.1.30}$$

It is not always possible to proceed in the opposite direction, i.e., to replace a system of n first order equations in n dependent variables by a single n'th order equation in one independent variable.

The transformation from (7.1.29) to (7.1.30) makes it possible to use a numerical method designed for solving a first order equation, to solve an n'th order equation.

For first order equations a number of theorems are known to have practical consequences. The Lipschitz condition plays a fundamental role in existence and uniqueness theorems.

Definition 7.1.9 *Let $f(x, \vec{y})$, where $\vec{y} = (y_1, \ldots, y_n)$, be a function of the $n+1$ variables $(x, \vec{y})$ in a convex domain D in $\Re^{n+1}$. We say that the function $f(x)$ satisfies Lipschitz conditions with respect to $\vec{y}$ in D if there exists a constant M (the Lipschitz constant) such that for all $(x, \vec{y})$ and $(x, \vec{z})$ in D*

$$|f(x, \vec{y}) - f(x, \vec{z})| \leq M \sum_{i=1}^{n} |y_i - z_i| \,, \tag{7.1.31}$$

where $\vec{z} = (z_1, \ldots, z_n)$.

In particular, the Lipschitz condition is satisfied if the partial derivatives of f with respect to y_i are all bounded in D; for then the moduli of the derivatives have a common upper bound M, that can serve as the Lipschitz constant.

Theorem 7.1.10 *Let the system of differential equations take the form*

$$y_i' = f_i(x, \vec{y}, \alpha, \beta, \ldots) \,, \quad i = 1, 2 \ldots, n$$

where $\alpha, \beta, \ldots$ are parameters. Then there exists a unique solution $y(x)$ that is a continuous vector function of the initial value $\vec{y}(x_0)$ and of the parameter $\alpha, \beta \ldots$ in every domain where the functions f_i are continuous functions of $(x, \vec{y}, \alpha, \beta, \ldots)$ and satisfy the Lipschitz condition with respect to y.

In this chapter our concern is mainly with a single first order equation, with a system of two first order equations, or with second order equations. Example 7.1.8 above illustrates that the analysis of a single first order equation can usually be extended to a system of n first order equations.

Problems

7.1.1 Find a power series solution of the problem

$$y'' = xy \ , \ \ y(0) = 1 \ , \ \ y'(0) = 0 \, .$$

7.1.2 Verify the inequality (7.1.23).

7.1.3 Let $Y(x)$ be a known solution of the differential equation

$$y'' + f(x)y' + g(x)y = 0 \ , \quad 0 \le x \le 1$$

where $Y(x) \neq 0$ in $[0,1]$. Assume that $u = vY$ and find a solution of

$$u'' + f(x)u' + g(x)u = h(x) \ , \quad 0 \le x \le 1 \ , \ \ u(0) = 1 \ , \ \ u'(0) = b \, .$$

7.1.4 Show that if $y = y(x)$ solves the problem

$$y'' + f(x,y) = 0 \ , \ \ 0 \le x \le 1 \ ; \quad y(0) = a \ , \ \ y(1) = b \, ,$$

then $y(x)$ also satisfies the equation

$$y(x) = a(1-x) + bx + \int_0^1 G(x,t) f(t, y(t)) dt$$

where

$$G(x,t) = \begin{cases} x(1-t) & \text{if } 0 \le t \le x \\ t(1-x) & \text{if } x \le t \le 1 \end{cases} .$$

Conversely, show that any twice differentiable function that satisfies (7.1.10) also satisfies the first equation given here.

7.1.5 Let M and N be two given functions of x and y such that $\frac{M}{y} = \frac{N}{x}$. Show that the differential equation

$$N(x,y)y' + M(x,y) = 0$$

has the "solution"

$$\int_{x_0}^{x} M(x, y_0) dx + \int_{y_0}^{y} N(x,t) dt = \text{const.}$$

7.1.6 Let m and k be the smallest positive integers such that $k = \frac{qm}{p+q}$, where p and q are integers, such that $q > 0$, $p + q > 0$. Show that the transformation $x = y^k$ reduces the integral $\int_0^1 x^{p/q} F(x)dx$ to the integral $\int_0^1 \frac{mq}{p+q}\, y^{m-1} F(y^k)dy$, i.e., it removes the singularity in the first integral if F has no singularities.

7.2 Methods for First Order Equations

Our goal in this section is two fold. With respect to the initial value problem

$$y' = f(x, y) \ , \ \ y(x_0) = y_0 \tag{7.2.1}$$

defined on the interval $[x_0, a]$, we outline the methods that are detailed in subsequent sections and second we explain the rationale behind these methods. We make an overriding assumption that a solution of the initial value problem on the interval $[x_0, a]$ exists and is unique. We denote this solution by $Y(x)$ where upper case notation distinguishes a true solution from an approximate.

We begin with the outline. In Section 7.3 we consider the Taylor series method applied to (7.2.1). As a practical tool, the Taylor series method is a reasonable choice only if (7.2.1) admits repeated differentiation. To make our point, the student of calculus need only think of computing the fifth derivative of the product of two functions. Clearly the circumstances are rare when Taylor series are useful other than as a theoretical tool.

We present difference methods in Section 7.4. These methods lead to approximations at the points

$$x_n = x_0 + n \ , \ \ n = 0, 1, 2, \ldots .$$

and, hence, are called discrete approximations. More specifically, y_n denotes the numerical estimate of the solution $Y(x_n)$. For notational purposes it is convenient to represent $Y(x_n)$ as Y_n and f_n as $f(x_n, y_n)$.

General k-step methods are the subject of Section 7.5. These methods take the form the form

$$y_{n+1} = \alpha_1 y_n + \alpha_2 y_{n-1} + \cdots + \alpha_k y_{n+1-k} + h(\beta_0 f_{n+1} + \beta_1 f_n + \cdots + \beta_k f_{n+1-k}) \,. \tag{7.2.2}$$

where $\alpha_1, \ldots, \alpha_k$ are coefficients whose specification distinguishes one k-step method from another. If the coefficients $\alpha_1, \ldots, \alpha_k$ and $\beta_0, \beta_1, \ldots, \beta_k$ in (7.2.2) are chosen such that

$$\gamma_0 \equiv \sum_{j=1}^{k} \alpha_j - 1 = 0 \ , \quad \gamma_1 \equiv \sum_{j=1}^{k} j\alpha_j - \sum_{j=0}^{k} \beta_j = 0 \tag{7.2.3}$$

and

$$\begin{aligned} \gamma_r &\equiv \sum_{j=0}^{k} \left[(1-j)^r \alpha_j + r(1-j)^{q-1} \beta_j \right] = 0 \ , \quad r = 2, \ldots, p \\ \alpha_0 &\equiv 1 \end{aligned} \tag{7.2.4}$$

then the method is said to have order p. The equations in (7.2.3) and (7.2.4) are obtained if we demand (7.2.2) to be satisfied exactly for $y(x) = x^r$, $r = 0, 1, \ldots, p$. Problem (7.2.1) shows that if the f in (7.2.1) is of class C^p in x and y and if $y_j = Y_j$ for $j = n+1-k, \ldots, n$ then

$$y_{n+1} - Y_{n+1} = \frac{\gamma_{p+1}}{(p+1)\,!} \, Y^{(p+1)}(\zeta) h^{p+1} \ , \quad x_{n+1-k} < \zeta < x_{n+1} \tag{7.2.5}$$

Explicit procedures for obtaining the constants α_j and β_j are explained in the following sections.

If $\beta_0 = 0$, then (7.2.2) is said to be explicit, since we can then explicitly solve for y_{n+1}, assuming that $y_n, y_{n-1}, \ldots, y_{n+1-k}$ are known. If $\beta_0 \neq 0$, then (7.2.2) is said to be implicit, since in that case y_{n+1} cannot in general be explicitly expressed. Specific forms of (7.2.2) are then used as correctors, for correcting the error of predictors (a predictor is almost always an explicit formula—see the examples in the next section) or else y_{n+1} is found via one of the methods of Chapter 3 for solving nonlinear equations. Except for the case of stiff problems (see Section 7.8), y_{n+1} is usually obtained via the following method of successive approximations:

$$\begin{aligned} y_{n+1}^{[0]} &= y_n \\ y_{n+1}^{[\nu+1]} &= h\beta_0 f\left(x_{n+1}, y_{n+1}^{[\nu]}\right) - u_n \ , \quad \nu = 0, 1, 2, \ldots \end{aligned} \tag{7.2.6}$$

where

$$u_n = \sum_{j=1}^{k} \left[\alpha_j y_{n+1-j} - h\beta_j f_{n+1-j} \right] . \tag{7.2.7}$$

If f is of class C^p in x and y and if h is sufficiently small then $y_{n+1}^{[\nu+1]} \to y_{n+1}$ as $\nu \to \infty$ and $y_{n+1}^{[\nu+1]}$ converges to the unique solution of (7.2.2) (see Problem 7.2.2). In practice, h is usually chosen sufficiently small so that no more than two or three iterations are required to compute y_{n+1} to the desired accuracy.

The integer k in (7.2.2) depends on the order p of the method. For larger orders p, it is necessary to take k larger in the construction of (7.2.2).

We also remark that the error expressed in (7.2.4) is proportional to $Y^{(p+1)}(\zeta)$ (of course, this is in general unknown) and h^{p+1}. In practice it is desirable that h be small so that this error may be small. However, if h is too small, the effect of roundoff error (see Section 7.13), the cost of computation as well as the effect of extraneous solutions (see Section 5.5) of (7.2.2) may destroy the accuracy of the computation. These extraneous solutions may grow more rapidly than the solution of (7.2.1) and mask it completely. Thus in practice we usually depend on crude criteria for obtaining a solution within a tolerable error, a criteria that is usually based on using (7.2.2) to compute an approximate solution for different values of h. Automatic programs exist that compute the step-size required to achieve a certain accuracy.

Equation (7.2.2) is simply a nonlinear finite difference equation of order k. Normally we do not attempt to obtain an explicit solution of this equation; instead we simply compute $y_k, y_{k+1}, \ldots$ and so on, repeating the process with a different value of h if necessary, until the error is acceptable. Even if we can explicitly solve the difference equation, we usually prefer a step-by-step solution for practical reasons.

Runge-Kutta methods are discussed in Section 7.6. The general Runge-Kutta method for solving (7.2.1) takes the form

$$y_1 = y_0 + h \sum_{j=0}^{n} a_j k_j \tag{7.2.8}$$

where

$$\begin{aligned} k_0 &= f(x_0 + \alpha_0 h, y_0) \\ k_1 &= f(x_0 + \alpha_1 h, y_0 + h\beta_{10} k_0) \\ k_2 &= f(x_0 + \alpha_2 h, y_0 + h\beta_{20} k_0 + h\beta_{21} k_1) \\ &\vdots \\ k_n &= f\left(x_0 + \alpha_n h, y_0 + \sum_{i=0}^{n-1} \beta_{ni} k_i\right) . \end{aligned} \tag{7.2.9}$$

The constants α_j, β_{ij} and a_j are chosen such that if y_0 is exact. As a result the sum on the right-hand side of (7.2.8) approximates the correct value of Y_1 to within an error of $0(h^p)$ as $h \to 0$, where p is as large as possible.

We also mention algorithms for computing an approximate solution of (7.2.1) using the notion of "extrapolation to the limit."

Euler's method

$$y_1 = y_0 + hf_0 \,,$$

is just the special case of (7.2.2) where we take $k = 1$, $\alpha_1 = 1$, $\beta_0 = 0$, is probably the simplest of all the methods for solving (7.2.1). It is also not very accurate, and its primary use therefore is to compute a rough approximation to the solution of (7.2.1).

Euler's method, the Taylor series method, Runge-Kutta methods, and methods based on extrapolation to the limit may be used to compute initial values for k-step methods, where $k > 1$.

Among methods in use, predictor-corrector methods based on Adams-type formulas and Runge-Kutta methods are the most popular. The predictor-corrector methods are self-checking, and they enable the user to obtain the solution within an acceptable tolerance. The popularity Runge-Kutta methods enjoy is because they are self-starting. Hence in contrast to predictor-corrector schemes, Runge-Kutta methods do not require an extra procedure to obtain the necessary starting values. We expect that due to their relative efficiency, Fehlberg's versions of Runge-Kutta methods will become popular during the next few years.

The usual method of estimating the error of Runge-Kutta methods recomputes the solution using the smaller step-size $h/2$. By making such a comparison at each step, we can determine the h required to achieve a certain tolerance of error.

Tests carried out by numerical analysts have shown that solution methods based on extrapolation to the limit can be very effective. The most popular of these methods is the method of Bulirsch and Stoer (see chapter references). Unfortunately this method is not yet fully understood. The main difficulty is that in contrast to the case of polynomial approximations, it is not always possible to tell a priori when a certain rational approximation scheme such as the Thiele algorithm (Problem 4.2.5) or the ϵ-algorithm produces a reliable approximation to a given function. Similar statements may be made regarding other methods of extrapolation to the limit.

It is hoped that the methods described in this chapter are some help to users faced with solving a problem of the type (7.2.1). In case of difficulty, we direct the user to more extensive literature(see chapter references).

Problems

7.2.1 Use the Taylor theorem expression

$$w(t) = \sum_{r=0}^{p} \frac{w^{(r)}(0)}{r\,!}\, t^r + \frac{1}{p\,!}\int_0^k (t-u)_+^p w^{(p+1)}(u)du$$

where $(t-u)_+ = \max(0, t-u)$ to prove that if (7.2.3) and (7.2.4) are satisfied, then

$$\sum_{s=0}^{k}[\alpha_s w(k-1) - \beta_s w'(k-s)] = \frac{\gamma_{p+1}}{(p+1)\,!}\, w^{(p+1)}(t)\,,, \tag{7.2.10}$$

for some $t \in (0, k)$. Use this result to verify (7.2.5).

7.2.2 Prove that if $|\beta_0|Mh < 1$, where M is a constant satisfying (1.28), then (7.2.2) has a unique solution y_{n+1}. Moreover, if $y_{n+1}^{[\nu]}$ is defined as in (7.2.6), then $y_{n+1}^{[\nu]} \to y_{n+1}$ as $v \to \infty$. Furthermore,

$$\left|y_{n+1}^{[\nu]} - y_{n+1}\right| \le \frac{(|\beta_0|Mh)^v}{1-|\beta|Mh}\left|y_{n+1}^{[1]} - y_n\right|\,.$$

7.3 Taylor Series Method of Solution of $y' = f(x, y)$

Let us assume that the partial derivatives of f with respect to x and y are easily obtained. We set

$$y(x) = \sum_{k=0}^{\infty} \frac{y^{(k)}(x_0)}{k\,!}\,(x-x_0)^k\,. \tag{7.3.1}$$

Equation (7.2.1) then yields

$$\begin{aligned}
y(x_0) &= y_0 \\
y'(x_0) &= f(x_0, y_0) \\
y''(x_0) &= f_x(x_0, y_0) + f_y(x_0, y_0)y'(x_0) \\
y^{(3)}(x_0) &= f_{xx}(x_0, y_0) + 2f_{xx}(x_0, y_0)y'(x_0) \\
&\quad + f_y(x_0, y_0)y''(x_0) + f_{yy}(x_0, y_0)y'(x_0)^2\,.
\end{aligned} \tag{7.3.2}$$

Substituting these values back into the series expansion (7.3.1), we get a polynomial approximation to the solution $Y(x)$ of (7.2.1).

If $f(x,y)$ is sufficiently simple, it is sometimes possible to get an explicit recurrence relation for the coefficients in (7.3.1) by means of the identity

$$\sum_{n=0}^{\infty} a_n x^n \sum_{m=0}^{\infty} b_m x^m = \sum_{n=0}^{\infty}\left(\sum_{m=0}^{n} a_{n-m} b_m\right) x^m \tag{7.3.3}$$

Example 7.3.1 Consider a series solution of $y' = x^2 + y^2$, $y(0) = 0$. We attempt the series

$$y = \sum_{n=0}^{\infty} c_n x^n . \tag{7.3.4}$$

Assuming that this series converges in some neighborhood of $x = 0$ we have

$$y' = \sum_{n=0}^{\infty}(n+1)c_{n+1}x^n \ , \quad y^2 = \sum_{n=0}^{\infty}\left(\sum_{m=0}^{n} c_m c_{n-m}\right) x^n .$$

Substituting these results into the differential equation, we arrive at the equation

$$\sum_{n=0}^{\infty}\left[(n+1)c_{n+1} - \delta_{2n} - \sum_{m=0}^{\infty} c_m c_{n-m}\right] x^n = 0 \,,$$

where $\delta_{2n} = 1$ if $n = 2$ and 0 otherwise. Hence we get the system of equation

$$c_0 = 0 \ , \quad c_{n+1} = \frac{1}{n+1}\left[\delta_{2n} + \sum_{m=0}^{\infty} c_m c_{n-m}\right] \ , \quad n = 0, 1, 2, \ldots .$$

Thus the coefficients of y_k are

$$\begin{aligned} c_0 &= c_1 = c_2 = 0 \ , \quad c_3 = \frac{1}{3} \\ c_4 &= c_5 = c_6 = 0 \ , \quad c_7 = \frac{1}{7}\, c_3^2 \ , \ldots \end{aligned}$$

These relations suggest that

$$c_{4k} = c_{4k+1} = c_{4k+2} = 0 \ , \quad c_{4k+3} \neq 0 \ , \quad k = 0, 1, 2, \ldots .$$

Indeed, this may be verified by induction. Putting

$$\begin{aligned} n+1 &= 4k+3 \to n = 4k+2 \\ m &= 4s+3 \to n-m = 4k-4s-1 \end{aligned}$$

it follows that

$$c_3 = \frac{1}{3} \ , \quad c_{4k+3} = \frac{1}{4k+3}\sum_{s=0}^{k-1} c_{4s+3}c_{4k-4s-1} \ , \quad k = 1, 2, 3, \ldots .$$

We use this recurrence relation to obtain the following values of c_{4k+3} given in Table 7.3.1a. These values are used in (7.3.4) to obtain approximates of $y' = x^2 + y^2$, $y(0) = 0$ in Table 7.3.1b.

Table 7.3.1: Results for Taylor Series Applied to Example 1.

(a)

k	c_{4k+3}
0	.3333 3333
1	.0158 7302
2	.0007 6200
3	.0000 5955
4	.0000 0370
5	.0000 0023
6	.0000 0001

(b)

x	y
0.0	0.0000 000
0.1	0.0003 333
0.2	0.0026 669
0.3	0.0090 035
0.4	0.0213 594
0.5	0.0417 912
0.6	0.0724 479
0.7	0.1156 598
0.8	0.1740 883
0.9	0.2509 097
1.0	0.3502 318

Problems

7.3.1 Find a cubic polynomial approximation to the problem $y'(x) = x^2 + y^2$, $y(0) = 0$ by computing (7.3.2) at $(x_0, y_0) = (0, 0)$. To find, for example, the next nonzero coefficient in (7.3.1) for this problem via continuing with the computation in (7.3.2) is quite tedious. The method of Example 7.3.1 is far simpler.

7.3.2 (a) Show that the problem

$$\begin{bmatrix} y' \\ v' \end{bmatrix} = \begin{bmatrix} v \\ y - \frac{y^3}{x^2} \end{bmatrix} \quad ; \quad \begin{bmatrix} y(0) \\ v(0) \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}$$

has a power series solution

$$\begin{bmatrix} y \\ v \end{bmatrix} = \begin{bmatrix} \sum_{n=0}^{\infty} a_n x^{2n+1} \\ \sum_{n=0}^{\infty} (2n+1) a_n x^{2n} \end{bmatrix}$$

where $a_0 = \alpha$ and

$$a_{n+1} = \frac{1}{(2n+2)(2n+1)} \left(\left[a_n - \sum_{j=0}^{n} \sum_{i=0}^{j} a_i a_{j-i} a_{n-j} \right] \right) \quad , \quad n = 0, 1, 2, \ldots .$$

(b) Prove that the expansions in (a) for the given coefficients converge if $|x| < \sqrt{6}/\sqrt{1+\alpha^2}$.

7.4 Finite Difference Approximations to $y' = f(x, y)$

In this section we derive some simple formulas of the type (7.2.2).

If we integrate both sides of (7.2.1) from x_r to x_m, we obtain

$$y_m - y_r = \int_{x_r}^{x_m} f(x, y)dx \, , \tag{7.4.1}$$

where $x_r = x_0 + rh$. For example, if we integrate from x_0 to x_1 the identity

$$Y'(x) = Y'(0) + (x - x_0)Y''(\zeta) \ , \quad x_0 < \zeta < x$$

and replace y_0' by f_0, we obtain the classical Euler formula

$$y_1 = y_0 + hf_0 \tag{7.4.2}$$

with error (see (7.2.5))

$$y_1 - Y_1 = \frac{h^2}{2} \, Y''(\zeta) \ , \quad x_0 < \zeta < x_1 \, . \tag{7.4.3}$$

The midpoint method is obtained by means of the identity

$$Y_{\pm 1} = Y(x_0 \pm h) = y_0 \pm hy_0' + \frac{h^2}{2} \, y_0'' \pm \frac{h^3}{6} \, y^{(3)}(\zeta_\pm)$$

where $\zeta_+(\zeta_-)$ is between x_0 and $x_1(x_{-1})$. By subtraction, we have

$$y_1 = y_{-1} + 2hf_0 \, . \tag{7.4.4}$$

The error is given by

$$y_1 - Y_1 = \frac{h^3}{3} \, Y^{(3)}(n) \ , \quad x_{-1} < n < x_1 \, . \tag{7.4.5}$$

Higher order approximations are derived by including more points. For example, if we interpolate f at x_{-2} and x_0 by a straight line and then integrate this approximation from x_{-2} to x_1, we obtain Heun's formula

$$y_1 - y_{-2} = \frac{3}{4} \, h(f_{-2} + 3f_0) \tag{7.4.6}$$

where the error is

$$y_1 - Y_1 = \frac{3}{8}\, h^4 Y^{(4)}(n)\ , \quad x_{-2} < n < x_1\,. \tag{7.4.7}$$

Similarly, if we intepolate f by a quadratic at x_{-2}, x_{-1} and x_0 and then integrate from x_{-2} to x_1, we obtain Milne's formula

$$y_1 = y_{-3} + \frac{4h}{3}\,(2f_{-2} - f_{-1} + 2f_0) \tag{7.4.8}$$

with error

$$y_1 - Y_1 = \frac{112}{360}\, h^5 Y^{(5)}(\zeta)\ , \quad x_{-3} < \zeta < x_1\,. \tag{7.4.9}$$

The rules (7.4.2), (7.4.4), (7.4.6) and (7.4.8) determine y_1 in terms of data at points to the left of x_1. In general, these relations are obtained by interpolating f to the left of x_1 with a Newton polynomial then replacing f in (7.4.1) by the Newton polynomial integrating . As may be expected by studying Chapter 4, these formulas are not very accurate. However, they do exhibit an advantage not shared by implicit rules; that is, formula (7.4.2) is self-starting hence requires no preliminary calculations. The formulas (7.4.4), (7.4.6) and (7.4.8) are not; for example, y_{-1} in (7.4.4) must be obtained prior to calculation.

Implicit formulas of the type (7.2.2) where $\beta_0 \neq 0$ are more generally more accurate than explicit formulas using the same points. To some extent this overcomes their starting difficulties. For instance, the formula

$$y_1 = y_0 + \frac{h}{2}\,(f_0 + f_1) \tag{7.4.10}$$

is more accurate than (7.4.2), since its error

$$y_1 - Y_1 = -\frac{h^3}{12}\, Y^{(3)}(\xi)\ , \quad x_0 < \xi < x_1 \tag{7.4.11}$$

exhibits a higher power of h. Of course, we caution that this new implicit procedure requires greater continuity. We refer to (7.4.10) as the trapezoidal rule. It is obtained by integrating from x_0 to x_1 the straight line that interpolates f at x_0 and at x_1. If f is a nonlinear function of y then we may not be able to solve (7.4.10) for y_1 directly, and it may be necessary to use the iterative methods given by (7.2.6) and (7.2.7). The equations (7.4.2) and

(7.4.10) may be used together as a predictor and corrector:

$$
\begin{aligned}
y_1^p &= y_0 + hf_0 \\
y_1^c &= y_0 + \frac{h}{2}\,(f_0 + f_1^p)\,,
\end{aligned}
\tag{7.4.12}
$$

where $f_1^p = f(x_1, y_1^p)$. This predictor-corrector pair is our first illustration of a *self-correcting formula*. If the difference $y_1^c - y_1^p$ is not within the tolerance of an acceptable error, the interval h must be decreased. The midpoint method (7.4.4) can be combined with the trapezoidal scheme (7.4.10) from x_{-1} to x_1 to yield the predictor-corrector scheme

$$
\begin{aligned}
y_1^p &= y_{-1} + 2hf_0 \\
y_1^c &= y_{-1} + h[f_{-1} + f_1^p]
\end{aligned}
\tag{7.4.13}
$$

Equations (7.4.5) and (7.4.11) show that

$$
\begin{aligned}
y_1^p - Y_1 &= \frac{h^3}{3}\,Y^{(3)}(\zeta) \\
y_1^c - Y_1 &= -\frac{h^3}{3}\,Y^{(3)}(\eta)
\end{aligned}
\tag{7.4.14}
$$

where ζ and η are between x_{-1} and x_1. The opposite sign of the error in these two formulas almost always makes the difference $y_1^p - y_1^c$ a reliable check on the error of computation.

Finally, the implicit formula

$$
y_1 = y_{-1} + \frac{h}{3}\,(f_{-1} + 4f_0 + f_1) \tag{7.4.15}
$$

with error

$$
y_1 - Y_1 = -\frac{h^5}{90}\,Y^{(5)}(n) \tag{7.4.16}
$$

may be used as a corrector for any of the procedures (7.4.2), (7.4.4), (7.4.6), or (7.4.8). The formula (7.4.15) is obtained by interpolating f at x_{-1}, x_0 and x_1 by a parabola and then integrating this parabola from x_{-1} to x_1.

The formula (7.4.15) is effective when used directly along with the iteration scheme (7.2.6)–(7.2.7). The result takes the form

$$\begin{aligned} y_1^{[0]} &= y_0 \\ y_1^{[v+1]} &= \frac{h}{3}\, f(x_1, y_1^{[v]}) + y_{-1} + \frac{h}{3}\,(4f_0 + f_{-1})\ , \quad v = 0, 1, 2, \ldots . \end{aligned} \tag{7.4.17}$$

A convenient property of implicit formulas such as (7.4.15) is that they are to some extent self-checking. For instance, if we suppose that (7.4.15) proceedes to a given point y_1 then we use (7.4.15) in the backward direction, i.e.,

$$y_{-1} = y_1 - \frac{h}{3}[f_{-1} + 4f_0 + f_1]$$

and verify the values already calculated. This practice is particularly useful in detecting unwanted growth of the solution (see Section 7.5).

Example 7.4.1 Solution of $y' = x^2 + y^2$, $y(0) = 0$, by Euler's method, $y_1 = y_0 + hf_0 = y_0 + h(x_0^2 + y_0^2)$. The results obtained are given in Table 7.4.1. Notice, upon comparing with the results of Table 7.3.1(b), that the error of the Euler method is quite large.

Table 7.4.1: Solution of $y' = x^2 + y^2$, $y(0) = 0$ by Euler's Method

x	y
0.0	0.0
0.1	0.0
0.2	0.001
0.3	0.0050 001
0.4	0.0140 026
0.5	0.0300 222
0.6	0.0551 123
0.7	0.0914 156
0.8	0.1412 518
0.9	0.2072 470
1.0	0.2925 421

Example 7.4.2 Solution of $y' = x^2 + y^2$, $y(0) = 0$, using the trapezoidal formula

$$y_1 = y_0 + \frac{h}{2}\left(x_0^2 + y_0^2 + x_1^2 + y_1^2\right).$$

Here y_1 can be conveniently solved for using the successive approximation technique of (7.2.6). The results obtained are given in Table 7.4.2. Notice that the results in Table 7.4.2 are considerably more accurate than those of the Euler method in Table 7.4.1.

Table 7.4.2: Solution of $y' = x^2 + y^2$, $y(0) = 0$ by the Trapezoidal Formula

x	y
0.0	0.0
0.1	0.0005 000
0.2	0.0030 005
0.3	0.0095 055
0.4	0.0220 343
0.5	0.0426 495
0.6	0.0735 106
0.7	0.1169 148
0.8	0.1756 922
0.9	0.2529 344
1.0	0.3528 586

Example 7.4.3 Solution of $y' = x^2 + y^2$, $y(0) = 0$, by the Milne-Simpson predictor-corrector method of (7.4.8) and (7.4.15):

$$\begin{aligned} y_1 &= y_{-3} + \frac{4h}{3}\left(2y'_{-2} - y'_{-1} + 2y'_0\right) \\ y_1 &= y_{-3} + \frac{2h}{3}\left(y'_{-3} + 4y'_{-1} + y'_1\right) \end{aligned}$$

The starting values at $x = 0.1, 0, .1$, and $.2$ are obtained from the series solution of Table 7.3.1.

In Table 7.4.3 the solution has been discontinued due to the large discrepancy between the results of the predictor and the corrector. In order to obtain more accuracy, a smaller mesh size should be used.

Table 7.4.3: Solution of $y' = x^2+y^2$, $y(0) = 0$ by the Milne-Simpson Method

x	y	y'
-.01	-.0003 333	.0100 001
0.0	.0000 000	.0000 000
0.1	.0003 333	.0100 001
0.2	.0026 669	.0400 071
0.3	.0090 019	Predictor
	.0090 055	Corrector
0.4	.0203 590	Predictor
	.0213 656	Corrector
0.5	.0418 046	Predictor
	.0417 794	Corrector

Problems

7.4.1 Compute a solution of the problem $y' = x^2+y^2$, $y(0) = 0$ by means of the method of (7.4.17) at $.2, .3, \ldots, 1.0$, given that $y_0 = 0$, $y_1 = 0.0003\,33333$. Compare the results with those in Table 7.3.1.

7.4.2 Using (7.2.2) write down an arbitrary $k = 1$ step method. Compare your result to

$$\alpha_1 = 1 \quad \text{and} \quad \alpha_1 - (\beta_0 + \beta_1) = 0$$

which is obtained for the more specific case (7.2.3).

7.4.3 (a) Note that explicit methods require $\beta_0 = 0$. This yields Euler's method (7.4.2).?? I don't understand this problem.

(b) Implicit methods allow $\beta_0 \neq 0$. Define all implicit one-step methods. To what selection does (7.4.10) correspond? Justify (7.4.11) by computing γ_2 in (7.2.4).

7.5 Adams Type Methods for Solving $y' = f(x,y)$

A vast number of methods for solving (7.2.1) can be obtained by operational methods. For example, beginning with (7.4.1) ($m = 1$, $r = 0$)

$$\begin{aligned} \int_{x_0}^{x_1} f(x)dx &= h\int_0^1 f(x_0 + ph)\, dp \\ &= \frac{-h\nabla}{(1-\nabla)\ln(1-\nabla)}\, f_0\,. \end{aligned} \tag{7.5.1}$$

On expanding in powers of the backward difference ∇, we obtain the a formal series

$$\int_{x_0}^{x_1} f(x)dx = h\sum_{\ell=0}^{\infty} a_\ell \nabla^\ell f_0 \tag{7.5.2}$$

that is exact for any polynomial f. In (7.5.2)

$$\begin{aligned} a_0 &= 1 \\ a_\ell &= 1 - \sum_{k=0}^{\ell-1} \frac{a_k}{k-\ell+1}\ , \quad \ell = 1, 2, 3, \ldots . \end{aligned} \tag{7.5.3}$$

In particular

$$a_0 = 1\ ,\ \ a_1 = \frac{1}{2}\ ,\ \ a_2 = \frac{5}{12}\ ,\ \ a_3 = \frac{3}{8}\ ,\ \ a_4 = \frac{251}{720}\,. \tag{7.5.4}$$

Since the backward differences of f_0 employ $f_0, f_{-1}, f_{-2}, \ldots$, i.e., the values of f to the left of f_0, the formulas obtainable by taking the first k terms of the expansion on the right-hand side of (7.5.2) are all explicit, and take the form

$$y_1 = y_0 + h\sum_{\ell=0}^{k} a_\ell \nabla^\ell f_0 \tag{7.5.5}$$

where the a_k are defined in (7.5.3). Moreover, if $y_0 = Y_0, y_{-1} = Y_{-1}, \ldots, y_{-k} = Y_{-k}$, and if $f \in C^{k+1}$, then (Problem 7.5.3)

$$y_1 - Y_1 = a_{k+1} h^{k+2} Y^{(k+2)}(\eta)\ , \quad x_{-k} < \zeta < x_1\,. \tag{7.5.6}$$

Similarly, we employ the formal representation (Problem)

$$\begin{aligned}\int_{x_0}^{x_1} f(x)dx &= -\frac{h\nabla}{\log(1-\nabla)}\, f_1 \\ &= h\sum_{\ell=0}^{\infty} b_\ell \nabla^\ell f_1\end{aligned} \tag{7.5.7}$$

that is exact for any polynomial f. In (7.5.7)

$$\begin{aligned} b_0 &= 1 \\ b_\ell &= -\sum_{k=0}^{\ell-1} \frac{b_k}{\ell-k+1}\ , \quad \ell = 1,2,3,\dots . \end{aligned} \tag{7.5.8}$$

In particular

$$b_0 = 1\ ,\ \ b_1 = -\frac{1}{2}\ ,\ \ b_2 = -\frac{1}{12}\ ,\ \ b_3 = -\frac{1}{24}\ ,\ \ b_4 = -\frac{19}{720}\,. \tag{7.5.9}$$

Truncation of the expansion (7.5.7) yields the implicit formula

$$y_1 = y_0 + h\sum_{\ell=0}^{k} b_1 \nabla^\ell f_1 \tag{7.5.10}$$

where the b_ℓ are defined as in (7.5.8) and where, in the notation of (7.5.6)

$$y_1 - Y_1 = b_{k+1} h^{k+2} Y^{(k+2)}(\eta)\ , \quad x_{1-k} < \zeta < x\ . \tag{7.5.11}$$

The equations (7.5.5) and (7.5.10) are often used together as predictor and corrector respectively. It can be seen by inspection that the initial coefficients b_ℓ are smaller in magnitude than the corresponding coefficients a_ℓ. Indeed, it can be shown that $b_\ell/a_\ell \to 0$ as $\ell \to \infty$, and that a_ℓ and b_ℓ are of opposite sign for all $\ell \geq 1$. Hence if (7.5.10) is used as a corrector, then for $Y^{(k+2)}(x)$ of constant sign on $[x_0, a]$, the errors in these formulas are opposite in sign. The methods defined by (7.5.5) and (7.5.10) yield some simple predictors and correctors in Tables 7.5.1 and 7.5.2.

Table 7.5.1: Adams Predictors

$$\begin{aligned} y_1 - y_0 &= h \sum_{\ell=0}^{k} a_\ell \nabla^\ell f_0 \\ a_0 = 1 \ , \ a_\ell &= 1 - \sum_{k=0}^{\ell-1} \frac{a_k}{\ell - k + 1} \end{aligned}$$

k	α_0^k	α_1^k	α_2^k	α_3^k	Error
0	1	0	0	0	$\frac{h^2}{2}\, Y''(\eta)$
1	3/2	-1/2	0	0	$\frac{5h^3}{12}\, Y'''(\eta)$
2	23/12	-16/12	5/12	0	$\frac{3h^4}{8}\, Y^{iv}(\eta)$
3	55/24	-59/24	37/24	-9/24	$\frac{251}{720}\, h^5 Y^{v}(\eta)$

$$y_1 - y_0 = h \sum_{\ell=0}^{k} \alpha_\ell^k f_{-\ell}$$

Problems

7.5.1 Use (a) the $k = 2$ and (b) the $k = 3$ predictor-corrector method of Table 7.5.1 to approximate $y(1)$, where $y' = y$, $y(0) = 1$, with $1/h = 4, 8, 16, 32$. For each of these cases, use the Taylor series approximation $y = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{120}$ to get the required initial values of y.

7.5.2 Verify the formula in (7.5.3). Hint: Recall the expansions

$$\frac{x}{1-x} = x^2 + x^3 + x^4 + \cdots$$

and

$$-\ell n(1-x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \cdots .$$

Put

$$\frac{x}{1-x} = \left(\sum_{j=0}^{\infty} a_j x^j \right) \left(\sum_{j=1}^{\infty} \frac{x^j}{j} \right)$$

and equate coefficients. (The same procedure verifies (7.5.9).)

Table 7.5.2: Adams Correctors

$$\begin{aligned} y_1 - y_0 &= h \sum_{\ell=0}^{k} b_\ell \nabla^\ell f_1 \\ b_0 = 1\ ,\ b_\ell &= -\sum_{k=0}^{\ell-1} \frac{b_k}{\ell - k + 1} \end{aligned}$$

k	β_0^k	β_1^k	β_2^k	β_3^k	Error
0	1	0	0	0	$\frac{-h^2}{2}\ Y''(\zeta)$
1	1/2	1/2	0	0	$\frac{-1}{12}\ h^3 Y'''(\zeta)$
2	5/12	8/12	-1/12	0	$\frac{-1}{24}\ h^4 Y^{iv}(\zeta)$
3	9/24	19/24	-5/24	1/24	$\frac{-19}{720}\ h^5 Y^{v}(\zeta)$

$$y_1 - y_0 = h \sum_{\ell=0}^{k} \beta_\ell^k f_{1-\ell}$$

7.5.3 Verify (7.5.6). Hint: Use the γ_r in (7.2.4).

7.6 Runge-Kutta Methods for Solving $y' = f(x, y)$

Let us refer to (7.2.9), and let us consider the case $n = 1$. We want to determine a_0, a_1, α_0, and α_1 so that

$$y_1 = y_0 + h\left[a_0 f(x_0 + \alpha_0 h, y_0) + a_1 f(x_0 + \alpha_1 h, y_0 + h\beta_{10} f(x_0 + \alpha_0 h, h_0))\right] \tag{7.6.1}$$

approximates $Y(x_0 + h)$ as accurately as possible in the limit as $h \to 0$. Using the notation

$$f^{(j,k)} = \frac{\partial^{j+k} f}{\partial^j x \partial^k y}$$

the function f, if sufficiently smooth, may be expanded as follows:

$$f(x_0 + u, y_0 + v) \quad = \quad \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \frac{f^{(j,k)}(x_0, y_0)}{j!\ k!}\ u^j v^k$$

$$= f(x_0,y_0) + uf_x(x_0,y_0) + vf_y(x_0,y_0) + 0(h^2) \tag{7.6.2}$$

as $h \to 0$, provided that both u and v are $0(h)$ as $h \to 0$. Also

$$Y(x_0+h) = Y(x_0) + hY'(x_0) + \frac{h^2}{2}\, Y''(x_0) + 0(h^3)$$

as $h \to 0$, and since $Y'(x) = f(x, Y(x))$, $Y(x_0) = y_0$,

$$Y(x_0+h) = y_0 + hf_0 + \frac{h^2}{2}\,[f_x(x_0,y_0) + f_0 f_y(x_0,y_0)] + 0(h^3) \tag{7.6.3}$$

as $h \to 0$.

Using (7.6.2) we get

$$\begin{aligned} k_0 &= f(x_0,y_0) + \alpha_0 h f_x(x_0,y_0) + 0(h^2)\ , \quad h \to 0 \\ k_1 &= f(x_0,y_0) + \alpha_1 h f_x(x_0,y_0) + \beta_{10} h k_0 f_y(x_0,y_0) + 0(h^2)\ , \quad h \to 0 \\ &= f(x_0,y_0) + \alpha_1 h f_x(x_0,y_0) + \beta_{10} h f(x_0,y_0) f_y(x_0,y_0) + 0(h^2)\ , \quad h \to 0 \end{aligned} \tag{7.6.4}$$

Substituting (7.6.4) into (7.6.1) we obtain

$$\begin{aligned} y_1 &= y_0 + (a_0 + a_1) h f_0 \\ &\quad + h^2\,[(a_0\alpha_0 + a_1\alpha_1) f_x(x_0,y_0) + a_1 \beta_{10} f_0 f_y(x_0,y_0)] \\ &\quad + 0(h^2)\ , \quad h \to 0 \end{aligned} \tag{7.6.5}$$

We now equate equal powers of h in (7.6.3) and (7.6.5) to arrive at the system of equations

$$\begin{aligned} a_0 + a_1 &= 1 \\ a_0\alpha_0 + a_1\alpha_1 &= \frac{1}{2} \\ a_1\beta_{10} &= \frac{1}{2} \end{aligned} \tag{7.6.6}$$

with one solution given by

$$\begin{aligned} a_0 &= a_1 = 1/2 \\ \alpha_0 &= 0\ , \ \ a_1 = 1 \\ \beta_{10} &= 1\,. \end{aligned} \tag{7.6.7}$$

This solution yields the Runge-Kutta method

$$y_1 = y_0 + \frac{h}{2}\,(k_0 + k_1) \tag{7.6.8}$$

where by (7.6.7),

$$\begin{aligned} k_0 &= f_0 \\ k_1 &= f(x_0 + h, y_0 + hk_0) \end{aligned} \tag{7.6.9}$$

and where if $y_0 = Y_0$, then $y_1 - Y_2 = 0(h^2)$, $h \to 0$. The method (7.6.8) with k_0 and k_1 as in (7.6.9) is just the previously derived predictor-corrector method (7.4.12). However, the system (7.6.2) has other solutions that yield Runge-Kutta methods that are different from (7.4.12). All of these are second order methods.

By means of the procedure used above, we can derive more accurate Runge-Kutta methods. The most popular of these is a fourth order method that takes the form

$$y_1 = y_0 + \frac{h}{6}\,(k_0 + 2k_1 + 2k_2 + k_3)\,, \tag{7.6.10}$$

where

$$\begin{aligned} k_0 &= f_0 \\ k_1 &= f\left(x_0 + \frac{1}{2}\,h\ ,\ y_0 + \frac{1}{2}\,hk_0\right) \\ k_2 &= f\left(x_0 + \frac{1}{2}\,h\ ,\ \ y_0 + \frac{1}{2}\,hk_1\right) \\ k_3 &= f(x_0 + h\ ,\ \ y_0 + hk_2) \end{aligned} \tag{7.6.11}$$

and where

$$y_1 - Y_1 = 0(h^5) \text{ as } \ h \to 0\,. \tag{7.6.12}$$

A very accurate variant of a Runge-Kutta method is due to Fehlberg. Given (7.2.1) one sets [34]

$$y = \bar{y} + y_0'(x - x_0) + \frac{1}{2}\,y_0''(x - x_0)^2 \tag{7.6.13}$$

where y_0' and $y_0''/2$ denote the Taylor coefficients of the solution y of (7.2.1), to get the problem

$$\bar{y}' = \bar{f}(x, \bar{y}) \ , \ \ \bar{y}(x_0) = \bar{y}_0 = y_0 \, , \tag{7.6.14}$$

where

$$\bar{f}(x, \bar{y}) = f(x, y) - y_0' - y_0''(x - x_0) \, .. \tag{7.6.15}$$

One next computes $\bar{y}_1$ by means of the formula

$$\bar{y}_1 = \bar{y}_0 + \frac{1}{1152}\,(500k_1 + 375k_2 + 64k_3) \tag{7.6.16}$$

where

$$\begin{aligned} k_1 &= h\bar{f}\left(x_0 + \frac{2}{5}\,h, \bar{y}_0\right) \\ k_2 &= h\bar{f}\left(x_0 + \frac{4}{5}\,h, \bar{y}_0 + \frac{4}{3}\,k_1\right) \\ k_3 &= h\bar{f}\left(x_0 + h, y_0 + \frac{5}{128}\,(27k_2 - 76k_1)\right) \, . \end{aligned} \tag{7.6.17}$$

The method is surprisingly accurate; it may be shown that

$$\bar{y}_1 - Y_1 = 0(h^7) \tag{7.6.18}$$

as $h \to 0$. The only drawback is that y_0' and y_0'' must be computed at each step. The computation of y_0' is trivial: $y_0' = f(x_0, y_0)$. Thus if y_0'' could also be easily computed at each step, the above method would be extremely efficient. Other more accurate variants of this procedure may be found in [35, 36, 37].

The actual error in (7.6.5) or (7.6.10) is difficult to estimate. Usually one applies a Runge-Kutta method at two different intervals over the same range and compares the results. A better approximation can then be obtained by extrapolation (Section 7.7).

Example: Runge-Kutta method of solution of $y' = x^2 + y^2$. $y(0) = 0$. The formula used is

$$y_1 = y_0 + \frac{h}{6}\,(k_0 + 2k_1 + 2k_2 + k_3)$$

where

$$\begin{aligned}
k_0 &= x_0^2 + y_0^2 \\
k_1 &= \left(x_0 + \frac{1}{2}\,h\right)^2 + \left(y_0 + \frac{1}{2}\,hk_0\right)^2 \\
k_2 &= \left(x_0 + \frac{1}{2}\,h\right)^2 + \left(y_0 + \frac{1}{2}\,hk_1\right)^2 \\
k_3 &= (x_0 + h)^2 + (y_0 + hk_2)^2 \,.
\end{aligned}$$

Using an interval $h = 0.2$, we obtain the solution given in Table 7.6.1.

Table 7.6.1: Solution of $y' = x^2 + y^2$, $y(0) = 0$, by a Runge-Kutta Method.

$h = .2$		$h = 1$	
x	y	x	y
0.0	0.0000 0000	0.1	.0003 333
0.2	0.0026 6687	0.2	.0026 669
0.4	0.0213 6009	0.3	.0090 035
0.6	0.0742 5120	0.4	.0213 594
0.8	0.1759 7804	0.5	.0417 913
1.0	0.3523 4874	0.6	.0724 481
		0.7	.1156 603
		0.8	.1740 810
		0.9	.2509 078
		1.0	.3502 337

The construction of fourth and higher order Runge-Kutta methods is very complicated. Nevertheless with sixth [19] and eighth [15, 20] order Runge-Kutta methods have been constructed. For example, finding the coefficients a_j, α_j, and β_{ij} for an eighth order Runge-Kutta method (7.2.8)–(7.2.9) requires the solution of 200 nonlinear algebraic equations.

Problems

7.6.1 Verify (7.6.12).

7.6.2 Show that when $f(x, y) = f(x)$ is independent of y, the fourth order Runge-Kutta method of Equations (7.6.10)–(7.6.11) reduces to Simpson's rule.

7.7 Extrapolation to the Limit

Extrapolation to the limit techniques may be used to improve the accuracy of an approximate solution of problem

$$y' = f(x, y) \qquad\qquad y(x_0) = y_0 \tag{7.7.1}$$

computed by a finite difference method of the type discussed in the previous sections. In this section we illustrate various procedures that may be used for carrying out extrapolation to the limit. All of the schemes we discuss involve obtaining an improved approximation to $g(0)$, where $g(h)$ is a function that can be computed for $h > 0$ and is known to have an asymptotic expansion of the form

$$g(h) \sim c_0 + c_1 h^{k_1} + c_2 h^{k_2} + c_3 h^{k_3} + \cdots \tag{7.7.2}$$

as $h \to 0$. In (7.7.2) the constants $c_0, c_1, c_2, \ldots$ are known to exist but their true values are not known, and $k_1 < k_2 < k_3 < \cdots$, are positive constants.

Indeed, under the absence of roundoff error, approximate solutions of (7.7.1) computed by means of one of the methods of the previous section do take the form (7.7.2) where c_0 denotes the exact solution. For example, the following result is established in [12, page 25].

Theorem 7.7.1 *Let f be of class C^p in x and y, and let $y(x, h)$ denote the solution of (7.7.1) computed by means of Euler's method, where $x_0 < x \le a$. Then*

$$y(x, h) = Y(x) + c_1(x)h + c_2(x)h^2 + \cdots + c_p(x)h^p + 0(h^{p+1}) \tag{7.7.3}$$

as $h \to 0$.

A similar result holds if $y(x, h)$ is computed by means of a higher order method. For example, for the case of a method of order 2, $c_1(x) = 0$, $c_2(x) \not\equiv 0$, unless $Y(x)$, the true solution of (7.7.1), is a parabola in x.

7.7.1 Richardson's Method of Extrapolation

Let us now describe the application of Richardson's method of extrapolation [17] to the function $g(h)$ that satisfies (7.7.2). The most rapid rate of convergence can be achieved in the case when the constants $k_1, k_2, k_3, \ldots$, are known.

We define a sequence $\{g_n(h)\}_{n=1}^{\infty}$ by

$$\begin{aligned} g_1(h) &= g(h) \\ g_{n+1}(h) &= g_n(h) + \frac{g_n(h) - g_n(qh)}{q^k n - 1} \end{aligned} \qquad (7.7.4)$$

where $q \neq 1$.

Theorem 7.7.2 *Let the sequence $\{g_n(h)\}_{n=1}^{\infty}$ be defined as above. Then there exist constants $c_n^{(n)}, c_{n+1}^{(n)}, c_{n+2}^{(n)}, \ldots$, such that*

$$g_n(h) \sim c_0 + c_n^{(n)} h^{k_n} + c_{n+1}^{(n)} h^{k_{n+1}} + c_{n+2}^{(n)} h^{k_{n+2}} + \cdots$$

as $h \to 0$.

Proof: The result (7.7.4) is clearly satisfied for the case $n = 1$. Assuming that the expansion (7.7.4) is valid for $n = m$, we find, upon substituting into (7.7.4), that

$$g_{m+1}(h) \sim c_0 + c_{m+1}^{(m+1)} h^{k_{m+1}} + c_{m+2}^{(m+1)} h^{k_{m+2}} + \cdots$$

as $h \to 0$, where

$$c_s^{(m+1)} = \frac{q^{k_m} - q^{k_s}}{q^{k_m} - 1}\ c_2^{(m)} \quad , \quad s = m+1, m+2, \ldots .$$

This completes the proof.

If the constants $k_1, k_2, \ldots$ are not known, the method of Equations (7.7.4) may still be applied by taking $k_n = n$, although the resulting sequence may not converge to the limit as rapidly in that case.

In practice one computes a table of entries $g_n(q^j h)$ where $0 < q < 1$ (usually $q = 1/2$) according to the layout of Table 7.7.1.

It then follows from Theorem 7.7.2 that each column of Table 7.7.1 converges to c_0, the rate of convergence of the n'th column being given by the asymptotic relation

$$g_n(q^j h) \sim c_0 + c_n^{(n)} (q^j h)^{k_n} + c_{n+1}^{(n)} (q^j h)^{k_{n+1}} + c_{n+2}^{(n)} (q^j h)^{k_{n+2}} + \cdots$$

as $q^j h \to 0$.

Table 7.7.1: Richardson's Method of Extrapolation

$g_1(h)$			
$g_1(qh)$	$g_2(h)$		
$g_1(q^2h)$	$g_2(qh)$	$g_3(h)$	
$g_1(q^3h)$	$g_2(q^2h)$	$g_3(qh)$	$g_4(h)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 7.7.2: Richardson's Extrapolation combined with Euler's Method

h	$y_1(1,h)$	$y_2(1,h)$	$y_3(1,h)$	$y_4(1,h)$
.2	.2428567456			
		.3422274636		
.1	.2925421046		.3497656855	
		.34788113		.3502140256
.05	.0320211617		.3501557983	
		.3495887698		
.025	.3349001935			

Example 7.7.3 Euler's Method and Richardson's Method of Extrapolation. We compute $y_1(1,h)$ (see Theorem 7.7.1) for the problem $y' = x^2+y^2$, $y(0) = 0$, by Euler's method for $h = 0.2, 0.1, .05$ and 0.025, to get the entries in column 1 of Table 7.7.2. Column 2 of this table is obtained by applying the formulas (7.7.4) with $k_1 = 1$, $q = 1/2$, while column 3 is obtained by applying (7.7.4) with $k_2 = 2$, $q = 1/2$.

7.7.2 Aitken's Method of Extrapolation

This method was previously discussed in Section 3.16. The layout again takes the form of Table 7.7.1, where, given $g_1(q^jh)$ we compute $g_2(q^jh), g_3(q^jh), \ldots$ by means of the formula

$$g_{n+1}(q^jh) = g_n(q^jh) - \frac{[g_n(q^{j+1}h) - q_n(q^jh)]^2}{g_n(q^{j+2}h) - 2g_n(g^{j+1}h) + g_n(q^jh)}. \tag{7.7.5}$$

Theorem 7.7.4 *Let $g_1(h)$ satisfy the relation (7.7.2) and let $0 < |q| < 1$. Then there exist constants $d_n^{(n)}, d_{n+1}^{(n)}, d_{n+3}^{(n)}, \ldots$, and $K_n^{(n)}, K_{n+1}^{(n)}, K_{n+2}^{(n)}, \ldots$, such that $K_1^{(0)} = k_1$, $K_n^{(n)} = K_n^{(n-1)}$, $n = 1, 2, 3, \ldots$, and such that*

$$g_n(q^j h) \sim c_0 + d_n^{(n)}(q^j h) K_n^{(n)} + d_{n+1}^{(n)}(q^j h) K_{n+1}^{(n)} + d_{n+2}^{(n)}(q^j h) K_{n+2}^{(n)} + \cdots \quad (7.7.6)$$

as $q^j h \to 0$.

We leave the proof of Theorem 7.7.4 to the problems.

Example 7.7.5 Euler's Method and Aitken's δ^2 Method. Starting with the values of $h_1(1,h)$ ($h = 0.2, 0.1, 0.05$) in Table 7.7.2, we compute an improved approximation $y_2^*(1,.2)$ to the value $Y(1)$ of the solution $Y(x)$ of the problem $y' = x^2 + y^2$, $y(0) = 0$, by means of (7.7.5):

$$\begin{aligned} Y(1) \simeq y_2(1,.2) &= y_1(1,.2) - \frac{[y_1(1,.2) - y_1(1,.1)]^2}{y_1(1,.2) - 2y_1(1,.1) + y_1(1,.05)} \\ &= .24285\ 67 - \frac{(0.5977\ 18)^2}{-.04218\ 87} \\ &= .32753\ 98\,. \end{aligned}$$

7.7.3 Extrapolation by Continued Fractions

The method of Bulirsch and Stoer uses a continued fraction procedure to extrapolate $g(0)$. We illustrate such a continued fraction method using the Thiele algorithm (Problem 4.2.5).

Example 7.7.6 Thiele's continued fraction method (Problem 4.2.5) yields

$$\begin{aligned} p(.2,.1) &= \frac{.2 - .1}{.24285\ 67 - .30262\ 85} = -1.6730\ 30 \\ p(.1,.05) &= \frac{.1 - .05}{.30262\ 85 - .24285\ 67} = -2.8436\ 40 \\ p(.2,.1,.05) &= \frac{.2 - .05}{-1.6730\ 30 - (-2.8436\ 40)} + .30262\ 85 \\ &= .43076\ 68 \end{aligned}$$

Hence, from Figure ???4.2.1, we get the convergent

$$r(h) = .24285\ 67 + \frac{h - .2}{-1.6730\ 30} + \frac{h - .1}{.43076\ 68 - .24285\ 67}$$

so that

$$y_1(1,0) \simeq r(0) = .33433\ 35\,.$$

The above examples show that each of three methods yields a different improved value of the limit. The problems that follow show that if more values of $y(1, h)$ are computed by means of Euler's method, then each of the three methods of extrapolation may be used to get a very accurate approximation of $y(1, 0)$.

While each of the columns in Table 7.7.1 converges to $g(0)$ for both Richardson's and Aitken's methods, it is not possible to prove, in general, that the n'th convergent of a continued fraction method converges to $g(0)$ as $n \to \infty$.

Problems

7.7.1 Continue the procedure of Example 1 of this section, by first using Euler's method to compute $y_1(1, .025)$ and $y_1(1, .0125)$ and then computing $y_n(1, .025)$ and $y_n(1, .0125)$ for $n = 2, 3, 4$, and 5 by means of the extrapolation method of (7.7.3).

7.7.2 Use the five computed values of $y_1(1, h)$ referred to in Problem 7.7.1 to construct a Thiele rational approximation, and use this approximation to extrapolate $y(1, 0)$.

7.7.3 Using the values $y_1(1, h)$ referred to in Problem 7.7.1 obtain a new sequence of approximations $y_2^*(1, h)$, $h = .2, .1, .05$, by means of Aitken's method.

7.7.4 Prove Theorem 7.7.4.

7.7.5 Devise a method of polynomial extrapolation to the limit based on:

(a) Aitken's algorithm (Example 9, Section 3.2).

(b) Neville's algorithm (Problem 3.2.4).

Apply each procedure to the values $y_n(1, h)$, $h = .2/2^{k-1}$, $k = 1, 2, 3, 4, 5$, and compare the results.

7.7.6 Let a Runge-Kutta method be of order p, i.e., if $y_0 = Y(x_0)$, then $y_1 = Y(x_1) + c_1 h^{p+1} + c_2 h^{p+2} + \cdots$. Let y_1^* and y_1^{**} be approximations of $Y(x_1)$ computed by the method using step sizes h and $h/2$ respectively.

(a) Discuss the use of the estimate

$$|y_1^{**} - Y(x_1)| \simeq \frac{|y_1^{**} - y_1^*|}{2^p - 1}$$

in the process of computation.

(b) In what sense is the approximation

$$y_1^R = y_1^{**} + \frac{y_1^{**} - y_1^*}{2^p - 1}$$

an improved approximation of $Y(x_1)$?

7.8 Direct Methods for Second Order Differential Equations

In this section we derive some numerical methods for solving the equations

$$y'' = g(x, y, z) \quad , \quad z = y' \tag{7.8.1}$$

and

$$y'' + P(x)y' + Q(x)y = R(x)\,. \tag{7.8.2}$$

We repeat a remark made in Section 7.1, namely that these equations can be solved by the methods of Sections 7.1 to 7.5, by writing them in the form of two first order equations. For example, (7.8.1) can be written in the form

$$\begin{aligned} y' &= z \\ z' &= g(x, y, z)\,. \end{aligned}$$

We can then apply any of the single equations methods to each of these equations.

The operator expansions (see 5.1.15 and 5.1.13)

$$hD = 2\sinh^{-1}(\delta/2) = 7\sum_{n=0}^{\infty} \frac{(1/2)_n(-1)^n\delta^{2n+1}}{(2n+1)2^{2n+1}n\,!} \tag{7.8.3}$$

and

$$\mu^p = \left(1 + \frac{1}{4}\,\delta^2\right)^{p/2} = \sum_{n=0}^{\infty} \frac{(-p/2)_n(-1)^n\delta^{2n}}{2^{2n}n\,!} \tag{7.8.4}$$

enable us to get expressions for higher derivatives, namely

$$\begin{aligned} h^r D^r &= \left[\frac{2}{\delta}\,\sinh^{-1}\frac{\delta}{2}\right]^r \delta^r \\ &= \left[1 - \frac{r}{24}\,\delta^2 + \frac{5r^2 + 22r}{5760}\,\delta^4 \right. \\ &\qquad \left. - \frac{35r^3 + 462r^2 + 1528r}{2903\,040}\,\delta^6 + \cdots\right]\delta^r \end{aligned} \tag{7.8.5}$$

and

$$h^r D^r = \left[\frac{1}{\mu}\left(\frac{2}{\delta}\,\sinh^{-1}\frac{\delta}{2}\right)^r\right]\mu\delta^r$$

$$
\begin{aligned}
&= \left[1 - \frac{r+3}{24}\,\delta^2 + \frac{5r^2+52r+135}{5760}\,q^4 \right. \\
&\qquad \left. - \frac{35r^3+777r^2+5749r+14175}{2903\ 040}\,\delta^6 + \cdots\right]\mu\delta^r\,.
\end{aligned} \tag{7.8.6}
$$

In these expansions we may take $r = 0, \pm 1, \pm 2, \ldots$.

The expansions (7.8.5) and (7.8.6) may be applied to (7.8.1) to yield the following expansions:

$$
\begin{aligned}
\delta^2 y_0 &= \left[\frac{2}{\delta}\ \sinh^{-1}\frac{\delta}{2}\right]^{-2} h^2 D^2 y_0 = h^2\left[\frac{2}{\delta}\ \sinh^{-1}\frac{\delta}{2}\right]^{-2} g_0 \\
&= h^2\left[1 + \frac{\delta^2}{12} - \frac{\delta^4}{240} + \frac{31\delta^6}{60\ 480} - \cdots\right] g
\end{aligned} \tag{7.8.7}
$$

$$
\begin{aligned}
hz_0 &= \left[\frac{1}{\mu}\left(\frac{2}{\delta}\ \sinh^{-1}\frac{\delta}{2}\right)\right]\mu\delta y_0 \\
&= \frac{1}{2}\left[\delta - \frac{\delta^3}{6} + \frac{\delta^5}{30} - \frac{\delta^7}{140} + \cdots\right]\left(y_{1/2} + y_{-1/2}\right)
\end{aligned} \tag{7.8.8}
$$

$$
\begin{aligned}
2\mu\delta z_0 &= 2h\left[\mu\left(\frac{2}{\delta}\ \sinh^{-1}\frac{\delta}{2}\right)^{-1}\right] D^2 y_0 \\
&= 2h\left[1 + \frac{\delta^2}{6} - \frac{\delta^4}{180} + \frac{\delta^6}{1512} - \cdots\right] g_0
\end{aligned} \tag{7.8.9}
$$

$$
\begin{aligned}
2\mu\delta y_0 &= 2h\left[\mu\left(\frac{2}{\delta}\ \sinh^{-1}\frac{\delta}{2}\right)^{-1}\right] z_0 \\
&= 2h\left[1 + \frac{\delta^2}{6} - \frac{\delta^4}{180} + \frac{\delta^6}{1512} - \cdots\right] z_0\,.
\end{aligned} \tag{7.8.10}
$$

In the numerical solution of (7.8.1) these expansions are truncated, and a possible predictor-corrector computation proceeds as follows. Once y_{-1}, z_{-1}, y_0 and z_0^p are known, we predict y_1 by means of the expression

$$
y_1^p = 2y_0 - y_{-1} + h^2 g(x_0, y_0, z_0^p) \tag{7.8.11}
$$

which is obtained from (7.8.7). Next, we correct z_0 by means of the expression

$$hz_0 = \frac{1}{2}\left(y_1^p - y_{-1}\right) \tag{7.8.12}$$

obtainable from (7.8.8). The value of z_1 is now predicted from a truncated form of (7.8.9), namely

$$z_1^p = z_{-1} + 2hg(x_0, y_0, z_0)\,. \tag{7.8.13}$$

We can then correct y_1 by means of a truncated form of the expansion (7.8.10), namely

$$y_1 = y_{-1} + \frac{h}{3}\left(z_{-1} + 4z_0 + z_1^p\right), \tag{7.8.14}$$

and we can next compute $g(x_1, y_1, z_1^p)$.

The predictor-corrector procedure of (7.8.11)–(7.8.14) eliminates nonlinear implicit schemes and successive approximations. Such an implicit method is described as follows.

Once y_{-1}, z_{-1}, y_0 and z_0 are known we can obtain more accurate values of y_1 and z_1 by using more terms in the expansions (7.8.7) and (7.8.9). For if we discard fourth and higher powers of δ in (7.8.7) and (7.8.9) we arrive at the system of two nonlinear equations

$$\begin{aligned} y_1 &= \frac{h^2}{12}\, g(x_1, y_1, z_1) + u \\ z_1 &= \frac{h}{3}\, g(x_1, y_1, z_1) + v \end{aligned} \tag{7.8.15}$$

where

$$\begin{aligned} u &= 2y_0 - y_1 + \frac{h^2}{12}\left(10g_0 + g_{-1}\right) \\ v &= z_{-1} + \frac{h}{3}\left(4g_0 + g_{-1}\right) \end{aligned} \tag{7.8.16}$$

and where $g_k = g(x_k, y_k, z_k)$, $k = -1, 0$. The system (7.8.15) is now solved by method of successive approximations:

$$\begin{bmatrix} y_1^{[0]} \\ z_1^{[0]} \end{bmatrix} = \begin{bmatrix} y_0 \\ z_0 \end{bmatrix}$$

$$\begin{bmatrix} y_1^{[v+1]} \\ z_1^{[v+1]} \end{bmatrix} = \begin{bmatrix} \frac{h^2}{12} g\left(x_1, x_1^{[v]}, z_1^{[v]}\right) + u \\ \frac{h}{3} g\left(x_1, y_1^{[v]}, z_1^{[v]}\right) + v \end{bmatrix} \quad v = 0, 1, 2, \ldots \; . \tag{7.8.17}$$

It can then be shown (Problem 7.8.4) that if $y_k = Y_k$, $z_k = Z_k$ for $k = -1, 0$, where Y is a solution of (7.8.1) and $Z = Y'$, then the solution $(y_1, z_1)^T$ of (7.8.15) satisfies the relation

$$\begin{bmatrix} y_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} Y_1 + \frac{h^6}{240} Y^{(6)}(\zeta) \\ Z_1 + \frac{h^5}{90} Y^{(6)}(\eta) \end{bmatrix}, \quad \begin{matrix} x_{-1} < \zeta < x_1 \\ x_{-1} < \eta < x_1 \end{matrix} \; . \tag{7.8.18}$$

Let us next describe a procedure for solving the second order equation (7.8.2). It is possible to eliminate predictors in this case, since the equation is linear. Using the expansions (7.8.5) and (7.8.6) to replace y^2D^2 and hD respectively in Equation (7.8.2), we get the approximating equations

$$\begin{aligned} (1 \quad &+ hP_0/2)y_1 = (2 - h^2Q_0)y_0 - (1 - hP_0/2)y_{-1} + h^2R_0 \\ &+ \left[\left(\tfrac{\delta^4}{12} - \tfrac{\delta^6}{90} + \tfrac{\delta^8}{150} - \cdots\right) + hP_0\mu\left(\tfrac{\delta^3}{6} - \tfrac{\delta^5}{30} + \tfrac{\delta^7}{140} - \cdots\right)\right] y_0 \end{aligned} \tag{7.8.19}$$

where in using this equation, the $0(h^4)$ (as $h \to 0$) term in square brackets is dropped.

In the case when $P(x) \equiv 0$, in (7.8.2) we apply the operator $(1 + \delta^2/12)$ to each side of (7.8.19) to get the very accurate Numerov method

$$\begin{aligned} (1 + h^2Q_1/12)y_1 &= (2 - 5h^2Q_0/6)y_0 - (1 + h^2Q_{-1}/12)y_{-1} \\ &\quad + \frac{h^2}{12}\left(R_1 + 10R_0 + R_{-1}\right) \\ &\quad - \left[\frac{\delta^6}{240} - \frac{13\delta^8}{15120} + \cdots\right] y \, . \end{aligned} \tag{7.8.20}$$

The $0(h^6)$ (as $h \to 0$) term in square brackets is dropped when the method (7.8.20) is applied. We illustrate the accuracy of Numerov's method in Example ?? below.

Example 7.8.1 Series solution of $y'' + xy = 0$, $y(0) = 1$, $y'(0) = 0$. It is readily verified that this problem has the power series solution

$$y = \sum_{n=0}^{\infty} y_n x^{3n}$$

where

$$y_0 = 1\ ,\ y_{n+1} = -\frac{1}{(3n+2)(3n+3)}\ y_n\ ,\ n = 0, 1, 2, \dots\ .$$

The first five coefficients are $y_0 = 1$, $y_1 = -0.1666\ 6667$, $y_2 = 0.0055\ 5556$, $y_3 = -0.0000\ 7716$, $y_4 = 0.0000\ 0058$. These enable us to compute "exact" values of y given in Table 7.8.1.

Table 7.8.1: The Exact Solution of $y'' + xy = 0$, $y(0) = 1$, $y'(0) = 0$

x	y
0.0	1.00000000
0.1	0.99983334
0.2	0.99866702
0.3	0.99550405
0.4	0.98935607
0.5	0.97925332
0.6	0.96425842
0.7	0.94348383
0.8	0.91611271
0.9	0.88142272
1.0	0.83881231

Example 7.8.2 Solution of $y'' + xy = 0$, $y(0) = 1$, $y'(0) = 0$ by Numerov's method. Using Equation (7.8.20) the values y_k are computed by means of the recurrence relation

$$(1 + h^2 x_1/12)y_1 = (2 - 5h^2 x_0/6)y_0 - (1 + h^2 x_{-1}/12)y_{-1}$$

where $y(-0.1)$ is computed by means of the power series representation of Example 1. The numerical values of y_p are given in Table 7.8.2.

Problems

7.8.1 In the algorithm described in (7.8.10)–(7.8.13), let y_{-1}, z_{-1}, and x_0^p be the exact values of the solution Y and $Z = Y'$ of (7.8.1) at x_{-1} and x_0 respectively. Prove that if $g \in C^2$ then the values z_1^p and y_1 satisfy the relations $z_1^p = Z_1 + 0(h^3)$ and $y_1 = Y_1 + 0(h^3)$ as $h \to 0$.

Table 7.8.2: The Solution of $y'' + xy = 0$, $y(0) = 1$, $y'(0) = 0$ by Numerov's Method

x	y
-0.1	1.0001667
0.0	1.0000000
0.1	0.9998333
0.2	0.9986670
0.3	0.9955040
0.4	0.9853560
0.5	0.9792532
0.6	0.9642583
0.7	0.9434837
0.8	0.9161126

7.8.2 Let Y be the solution of (7.8.2) and let y_1 be determined from the (7.8.14), with the term in square brackets replaced by zero. Use the method of Problem 2.1 to prove that if P, Q, and R are of class C^2 then

$$y_1 - Y_1 = \frac{h^4}{12}\, Y^{(4)}(\zeta) - \frac{h^4}{6}\, P_0 Y^{(3)}(\eta)\,,$$

where ζ and η are between x_{-1} and x_1.

7.8.3 In (7.8.2), let $P(x) = 0$, and let Y be a solution of this equation. Use the method of Problem 2.1 to show that if $y_{-1} = Y_{-1}$, $y_0 = Y_0$, and if y_1 is determined by means of (7.8.20) with the term in square brackets replaced by zero, then

$$y_1 - Y_1 = [1 + h^2 Q_1/12]^{-1} Y^{(6)}(\zeta) \quad , \quad \text{for some } \zeta \in (x_{-1} x_1)\,.$$

7.8.4 Prove the result of (7.8.18).

7.8.5 $*$ Nystrom developed a Runge-Kutta-type method that is especially suited to the problem

$$y'' = f(y)\ ,\ y(x_0) = y_0\ ,\ y'(x_0) = z(x_0) = y_0'\,.$$

Applied to this problem Nystrom's [21] method is

$$\begin{aligned} z_{k+1} &= z_k + \frac{h}{192}(23k_0 + 125k_1 - 81k_2 + 125k_3) \\ y_{k+1} &= y_k + hz_k + \frac{h^2}{192}(23k_0 + 75k_1 - 27k_2 + 25k_3) . \end{aligned}$$

Here

$$\begin{aligned} k_0 &= f(y_k) \\ k_1 &= f\left(y_k + \tfrac{2}{5}\, hz_k + \tfrac{2}{25}\, h^2 k_0\right) \\ k_2 &= f\left(y_k + \tfrac{2}{3}\, hz_k + \tfrac{2}{9}\, h^2 k_0\right) \\ k_3 &= f\left(y_k + \tfrac{4}{5}\, hz_k + \tfrac{4}{25}\, h^2 (k_0 + k_1)\right) . \end{aligned}$$

Show that

$$y_1 - Y(x_1) = 0(h^5) \quad \text{as} \quad h \to 0 .$$

7.9 Boundary Value Problems

Let us consider the numerical solution of the boundary value problem

$$y'' = g(x, y, y') \ , \quad y(0) = y(1) = 0 . \tag{7.9.1}$$

We remark that if a and b are finite, the problem

$$\frac{d^2w}{dt^2} = g(t, w, w') \ , \quad w(a) = A \ , \quad w(b) = B \tag{7.9.2}$$

can be transformed into the problem (7.9.1) by means of the transformations

$$\begin{aligned} t &= a + (b-a)x \\ w &= y + A(1-x) + Bx . \end{aligned} \tag{7.9.3}$$

Not all problems of the type (7.9.1) have a solution, nor is it always possible to tell, in practice, whether or not a particular problem of the type (7.9.1) has a solution. Sometimes it is, however, possible to give sufficient conditions so that a class of problems of the type (7.9.1) have a solution. For example, if $g(x, y, y') = G(x, y)$ where $G_y(x, y) \geq 0$ for all $x \in [0, 1]$ and all real y, then the problem (7.9.1) has a unique solution. Let us illustrate the application of matrix methods to some linear boundary value problems.

Example 7.9.1 Matrix solution of the boundary problem $y'' + y + 1 = 0$, $y(0) = y(1) = 0$. We use the approximation $h^2 y'' = \delta^2 y$. The exact solution of the boundary value problem is $y = \{[1 - \cos(1)]/\sin(1)\} \sin x + \cos x - 1$. We take $h = 0.1$, and take account of symmetry about $x = 0.5$, to get the system of equations (see Section 5.5.2)

$$\begin{bmatrix} 1.99 & -1 & & & \\ -1 & 1.99 & -1 & & \\ & -1 & 1.99 & -1 & \\ & & -1 & 1.99 & -1 \\ & & & -2 & 1.99 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 0.01 \\ 0.01 \\ 0.01 \\ 0.01 \\ 0.01 \end{bmatrix}$$

which has the solution given in Table 7.9.1.

Table 7.9.1: The solution of $y'' + y + 1 = 0$, $y(0) = y(1) = 0$

Approximate	Exact
$y_1 = y_9 = 0.04958$	0.04954
$y_2 = y_8 = 0.08868$	0.08860
$y_3 = y_7 = 0.11688$	0.11678
$y_4 = y_6 = 0.13392$	0.13380
$y_5 = 0.13962$	0.13949

Example 7.9.2 Solution of $y''+y+1 = 0$, $y(0) = y(1) = 0$ by computing two solutions f_p and g_p and taking $y_p = \alpha f_p + (1 - \alpha) g_p$ (see Section 5.5.2).

We again take $h = 0.1$ and use the difference scheme

$$y_{p+1} = 1.99 y_p - y_{p-1} - 0.01 \quad , \quad p = 1, 2, 3, \ldots$$

as in the previous example. We compute f_p such that $f_0 = f_1 = 0$ and g_p such that $g_0 = 0$, $g_1 = 1$. After computing the two solutions, we compute α such that $y_{10} = \alpha f_{10} + (1 - \alpha) g_{10} = 0$. The results are contained in Table 7.9.2.

Table 7.9.2: Solution of $y'' + y + 1 = 0$, $y(0) = y(1) = 0$

$1 - \alpha = 0.04958$

x	f	g	y
0.0	0.00000	0.00000	0.00000
0.1	0.00000	1.00000	0.04958
0.2	−0.01000	1.98000	0.08868
0.3	−0.01990	2.93020	0.11688
0.4	−0.05950	3.84109	0.13392
0.5	−0.09850	4.70358	0.13962
0.6	−0.14652	5.50904	0.13392
0.7	−0.20308	6.24940	0.11688
0.8	−0.26760	6.91728	0.08868
0.9	−0.33945	7.50598	0.04959
1.0	−0.49791	8.00962	0.00000

Example 3: An eigenvalue problem, $y'' + \theta xy = 0$, $y(0) = y(1) = 0$.

(a) Exact solution by power series method. The solution of the initial value problem

$$y'' + xy = 0\ ,\ \ y(0) = 0\ ,\ \ y'(0) = 1$$

is

$$y(x) = x - \frac{1}{12}\, x^4 + \frac{1}{6 \cdot 12^2}\, x^7 - \cdots$$

at $x = 1$, we must have

$$y(1) = 1 - \frac{1}{12} + \frac{1}{6 \cdot 12^2} - \cdots\ .$$

The exact value of θ_1 to seven significant figures, is found using successive approximations, and is $\theta_1 = 18.95622$.

(b) Approximate solution via a system of algebraic equations is carried out by approximation $h^2 y_p''$ by $\delta^2 y_p$. This leads to the following system of homogeneous equations:

$$\begin{bmatrix} 2-\mu & -1 & & & \\ -1 & 2-2\mu & -1 & & \\ & -1 & 2-3\mu & -1 & \\ & & \cdots & \cdots & \\ & & & -1 & 2-n\mu \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where $h = 1/n$, $\mu = \theta h^3$. Setting the determinant of this system equal to zero, we obtain the following results.

$$\begin{array}{lcl}
n=2: & \mu - 2 = 0 & \theta_1 = 16 \\
n=3: & 2\mu^3 - 6\mu + 3 = 0 & \\
 & \mu_1 = 0.633974 & \theta_1 = 17.11732 \\
 & \mu_2 = 2.366025 & \theta_2 = 63.88268 \\
n=4: & 6\mu^3 - 22\mu^2 + 20\mu - 4 = 0 & \\
 & \mu_1 = 0.27924075 & \theta_1 = 17.87141 \\
 & \mu_2 = 1.0 & \theta_2 = 64.0 \\
 & \mu_3 = 2.587426 & \theta_3 = 152.79526 \\
n=5: & 24\mu^4 - 100\mu^3 + 122\mu^2 - 50\mu + 5 = 0 & \\
 & \mu_1 = 0.1460119 & \theta_1 = 18.2510\,49
\end{array}$$

Using Aitken's δ^2 process on the above values of θ_1, we get the improved approximation $\theta_1 = 18.92\,719$.

Problems

7.9.1 Let a and b be finite, and consider the second order differential equation problem

$$\begin{aligned} y'' + f(x)y' \quad + g(x)y &= r(x) \quad , \quad a \le x \le b \\ y(a) = A \qquad\qquad & , \quad y(b) = B \end{aligned}$$

(a) Replace h^2y'' by $\delta^2 y$, hy' by $\mu\delta y$, where $h = (b-a)/n$ to obtain a linear system with tridiagonal matrix U.

(b) Show that if f and g are uniformly bounded on $[a, b]$ and if $g(x) < 0$ on $[a, b]$ then the matrix U is non-singular for all h sufficiently small.

7.10 The Shooting Method of Solving Boundary Value Problems

We describe two variants of the shooting method. If we denote the solution of the initial value problem

$$y'' = g(x, y, y') \ , \ y(0) = 0 \ , \ y'(0) = \alpha \tag{7.10.1}$$

by $Y(x, \alpha)$ then we want to find α such that $Y(1) = Y(1, \alpha) = 0$. We may thus solve for $Y(1, \alpha)$ by the Runge-Kutta method and then use an iterative method of solving the nonlinear equation $Y(1, \alpha) = 0$. For Newton's method we also require $W(1) = W(1, \alpha)$ where $W(x, \alpha) = \frac{\partial y(x,\alpha)}{\partial \alpha}$, and where the function $W(x) = W(x, \alpha)$ is the solution of the initial value problem

$$\begin{aligned} W'' &= g_y(x, Y(x), Y'(x))W + g_{y'}(x, Y(x), Y'(x))W' \\ W(0) &= 0, W'(0) = 1 \, . \end{aligned} \tag{7.10.2}$$

Let us describe another variant of the shooting method. Suppose that we approximate the differential equation (7.10.1) by the difference equations

$$y_{k+1} = y_k + F(y_{k-1}) \tag{7.10.3}$$

where $y_0 = 0$, $y_1 = a$. We want to find a such that $y_n = y_n(a) = 0$. This can be done using Newton's method, provided that along with Equations (7.10.3) we also solve the equation

$$w_{k+1} = w_k + F'(y_{k-1})w_{k-1} \ ; \ w_0 = 0, w_1 = 1 \, . \tag{7.10.4}$$

Here F' denotes the derivative of F, and $w_k = dy_k/da$. Hence with $a = a_0$, we simultaneously compute y_k and w_k, $k = 2, 3, \ldots, n$, and then get an improved value of a using the Newton formula

$$a_1 = a_0 - \frac{y_n(a_0)}{w_n(a_0)} \, . \tag{7.10.5}$$

We then recompute y_n and w_n with a_0 replaced by a_1, and repeat the process until a desired accuracy is achieved.

The shooting method can also be used to compute eigenvalues. For example, we may wish to find a value of the parameter c such that the problem

$$y'' + g(x, c)y = 0 \ , \ y(0) = y(1) = 0 \tag{7.10.6}$$

has a non-trivial solution. Here we can approximate (7.10.6) by

$$y_{k+1} - 2y_k + y_{k-1} + h^2 g(x_k, c) y_k = 0 \tag{7.10.7}$$

where we take $y_0 = 0$, $y_1 = h$. Along with (7.10.7) we also solve

$$\begin{aligned} w_{k+1} \quad & -2w_k + w_{k-1} + h^2 g_c(x_k, c) y_k + h^2 g(x_k, c) w_k = 0 \\ w_0 \quad & = w_1 = 0 \end{aligned} \tag{7.10.8}$$

where $Y_k = dy_k/dc$. Thus starting with $c = c_0$, we get an improved value c_1 of c by means of the formula

$$c_1 = c_0 - \frac{y_n(c_0)}{w_n(c_0)} \,. \tag{7.10.9}$$

We then replace c_0 by c_1 and recompute y_n, w_n, and c_1, and so on, until a desired accuracy is achieved.

Example 7.10.1 The problem

$$\begin{aligned} \frac{d^2y}{dx^2} - \left(\frac{1}{U-c} + \frac{1}{U-c-c_r}\right) U' \frac{dy}{dx} - A^2 \left(\frac{U-c-c_r}{U-c}\right) y \quad &= 0 \\ y(0) = y(1) \quad &= 0 \end{aligned}$$

arises in weather forecasting. Here y is related to the vertical wind velocity, A and c_r are given constants and $U = U(x)$ is a given function. The problem is to determine the value of c such that a non-trivial solution exists. For the case of $U = 10x(1-x)^3$, $c_r = 2.14$, we replace y_p'' by $(y_{p+1} - 2y_p + y_{p-1})/h^2$, y_p' by $(y_{p+1} - y_{p-1})/(2h)$, take $h = 1/36$ and apply the above described Newton's method to solve for two values of c for $A = 0, 1, 2$. We thus find that

$$\begin{aligned} c \quad &= 0.43030\,, \quad -1.56988 \ \text{ with } \ A = 0 \\ c \quad &= 0.36213\,, \quad -1.57459 \ \text{ with } \ A = 1 \\ c \quad &= 0.37474\,, \quad -1.61654 \ \text{ with } \ A = 2\,. \end{aligned}$$

We can obtain an exact value of c in the case where $A = 0$, since the differential equation can be explicitly solved then. The exact values of c when $A = 0$ are $c = 0.42904$ and $c = -1.56904$.

Problems

7.10.1 Let the solution of Problem 7.9.1 be approximated by the solution of the difference equations (7.10.7), so that y_n is computed as a function of y_1. Thus we want to solve the nonlinear equation $(*)$ $y_n(y_1) = 0$. Construct an algorithm for solving the problem $(*)$ by combining the method of (7.10.3) with the secant method.

7.11 Solution of Linear Second Order Differential Equations using Splines

Let us subdivide the interval $[a, b]$ into n equal intervals $[x_{j-1}, x_j]$, $j = 1, 2, \ldots, n$, where $x_0 = a$, $x_j = x_0 + jh$, and where $x_n = x_0 + nh = b$. Let us interpolate a function $y(x)$ on the interval $[a, b]$ by $S(x)$ where $S(x)$ is a cubic polynomial in each of the intervals $[x_{j-1}, x_j]$, such that $S(x_j) = y_j$ and such that $S(x)$ and $S'(x)$ are continuous on $[a, b]$. $S(x)$ is then called a cubic spline. It may be readily verified that the solution is given on the interval $[x_{j-1}, x_j]$ by the equation

$$\begin{aligned} S(x) \quad &= \quad M_{j-1}\frac{(x_j - x)^3}{6h} + M_j\frac{(x - x_{j-1})^3}{6h} \\ &\quad + \left(y_{j-1} - \frac{h^2}{6}M_{j-1}\right)\frac{(x_j - x)}{h} \\ &\quad + \left(y_j - \frac{h^2}{6}M_j\right)\frac{(x - x_{j-1})}{h} \end{aligned} \tag{7.11.1}$$

where $M_j = S''(x_j)$ and $y_j = y(x_j)$. Hence for $j = 0, 1, \ldots, n-1$,

$$S'(x_j^+) = -\frac{h}{3}M_j - \frac{h}{6}M_{j+1} + \frac{y_{j+1} - y_j}{h} \tag{7.11.2}$$

while for $j = 1, 2, \ldots, n$,

$$S'(x_j^-) = \frac{h}{3}M_j + \frac{h}{6}M_{j-1} + \frac{y_j - y_{j-1}}{h}\,. \tag{7.11.3}$$

Continuity of $S'(x)$ on $[a, b]$ therefore implies that $S'(x_j^+) = S'(x_j^-)$, i.e. that

$$\frac{h}{6}M_{j-1} + \frac{2h}{3}M_j + \frac{h}{6}M_{j+1} = \frac{y_{j+1} - 2y_j + y_{j-1}}{h}\ , \quad j = 1, 2, \ldots, n-1\,. \tag{7.11.4}$$

Given the boundary value problem

$$\begin{aligned} y'' + P(x)y' + Q(x)y \quad &= \quad R(x)\ , \quad a \le x \le b \\ y(a) = A\ ,\ \ y(b) \quad &= \quad B\,, \end{aligned} \tag{7.11.5}$$

we can approximate the solution $y(x)$ using the above cubic spline by making the spline approximation satisy the differential equation (7.11.5) at the *knots*

x_j, $j = 0, 1, \ldots, n$. Using (7.11.2) and (7.11.3) this leads to a set of relations from which we can eliminate the unknowns $M_0, M_1, \ldots, M_n$. The result, in conjunction with the boundary conditions in (7.11.5), is a set of tridiagonal equations for the determination of $y_0, y_1, \ldots, y_n$.

For the case when $P(x) \equiv 0$ the differential equation yields

$$M_j = R_j - Q_j y_j \ , \quad j = 0, 1, \ldots, n \tag{7.11.6}$$

that when substituted into (7.11.4) yields

$$\begin{aligned} y_{j+1}\left(1 + \frac{h^2}{6}Q_{j+1}\right) & -y_j\left(2 - \tfrac{2h^2}{3}Q_j\right) + y_{j-1}\left(1 + \tfrac{h^2}{6}Q_{j-1}\right) \\ & = \tfrac{h^2}{6}\left(R_{j+1} + 4R_j + R_{j-1}\right) \ , \quad j = 1, 2, \ldots, n-1 \ . \end{aligned} \tag{7.11.7}$$

Using differences we may write this equation as

$$\delta^2 y_j + h^2\left(1 + \frac{1}{6}\delta^2\right)Q_j y_j = h^2\left(1 + \frac{1}{6}\delta^2\right)R_j \ . \tag{7.11.8}$$

Thus (7.11.7) has a local truncation error of order h^4 as $h \to 0$. The equation (7.11.7) is also similar to the Numerov formula (7.8.20) except that $\delta^2/6$ now replaces $\delta^2/12$ in (7.8.20).

Cubic splines possess a curious approximation property.

Theorem 7.11.1 *Let $f \in C^2[a, b]$, and let $S(x)$ by the cubic spline that satisfies*

$$\begin{aligned} S(x_i) &= f(x_i) \ , \quad i = 0, 1, \ldots, n \\ S'(a) &= f'(a) \ , \quad S'(b) = f'(b) \ . \end{aligned}$$

Then

$$\int_a^b [f''(x) - S''(x)]^2 dx \le \int_a^b [f''(x) - V''(x)]^2 dx \tag{7.11.9}$$

where V is any function that reduces to a polynomial of degree at most 3 on each interval $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, n$.

Example: Solution of $y'' + y + 1 = 0$, $y(0) = y(1) = 0$ by use of cubic splines. For the sake of illustration, we divide the interval $[0, 1]$ into two subintervals, $[0, 0.5]$ and $[0.5, 1]$. Then using (7.11.7), we get

$$\begin{aligned} y(0.5) &= \frac{1}{2 - \frac{2(0.5)^2}{3}} \frac{(0.5)^2}{6} (1 + 4 + 1) \\ &= \frac{3}{22} \simeq 0.1364 \,. \end{aligned}$$

Next, using (7.11.6), we get

$$M_0 = M_2 = -1 \;, \;\; M_1 = -\frac{47}{48} \,.$$

Substituting these results into (7.11.1), we get

$$S(x) = \frac{47}{48} x - \frac{1}{2} x^2 - \frac{1}{22} x^3 + \frac{u}{11} \left(x - \frac{1}{2} \right)^3 \,,$$

where $u = 0$ for $x < 0.5$, $u = 1$ for $x \geq 0.5$.

We cannot expect much accuracy using such a large interval. For example, for $x = .3$, our solution yields $S(x) = S(.3) \simeq .2475$, whereas from Table 7.9.1, $y(.3) \simeq .1168$.

The case when the first derivative is present in (7.11.5), i.e. when $P(x) \not\equiv 0$, leads upon the elimination of the M_j, to a more complicated difference equation, namely

$$\begin{aligned} y_{j+1} \left(1 + \frac{h}{2} P_{j+1} + \frac{h^2}{6} Q_{j+1} \right) A_j - y_j \left[\left(1 + \frac{h}{2} P_{j+1} \right) A_j + \left(1 - \frac{h}{2} P_{j-1} \right) B_j - \frac{2h^2}{3} Q_j C_j \right] \\ + y_{j-1} \left(1 - \tfrac{h}{2} P_{j-1} + \tfrac{h^2}{6} Q_{j-1} \right) B_j \\ = \tfrac{h^2}{6} \left(A_j R_{j+1} + 4 C_j R_j + B_j R_{j-1} \right) \,, \;\; j = 1, 2, \ldots, n-1 \end{aligned} \tag{7.11.10}$$

where

$$\begin{aligned} A_j &= \left(1 - \tfrac{h}{3} P_{j-1} \right) \left(1 + \tfrac{h}{3} P_j \right) + \tfrac{h^2}{36} P_{j-1} P_j B_j \quad = A_{j+1} \\ C_j &= 1 + \tfrac{7h}{24} \left(P_{j+1} - P_j \right) - \tfrac{h^2}{12} P_{j-1} P_{j+1} \,. \end{aligned}$$

Problems

7.11.1 Derive (7.11.1) by starting with the approximation

$$f''(x) \simeq S''(x) = M_{j-1}\frac{x - x_{j-1}}{h} + M_j\frac{(x_j - x)}{h}\,.$$

7.11.2 * Verify the relation (7.11.9).

7.12 The Bubnov-Galerkin Method for Solving Second Order Boundary Value Problems

Let us consider the problem

$$y'' + f(x, y, y') = 0\ ,\ \ y(0) = y(1) = 0\,. \tag{7.12.1}$$

We assume that the problem (7.12.1) has a solution $y_0(x)$, such that for some $\delta > 0$ there is no other solution $y(x)$ in the region $S_\delta(y_0) = \{y : \|y - y_0\| < \delta\}$. Here we adopt the notation

$$(u, v) = \int_0^1 u(x)v(x)dx \tag{7.12.2}$$

and $\|u\| = (u, u)^{1/2}$. Furthermore, we assume that f can be differentiated at least once with respect to each of the variables x, y, and y', for all y and y' such that $\|y - y_0\| + \|y' - y_0'\| < \delta$.

The Bubnov-Galerkin method of obtaining an approximate solution y_n of (7.12.1) is to set

$$y_n = \sum_{k=1}^{n} a_k u_k\ ,\ \ u_k(x) = 2^{1/2}\sin k\pi x\,, \tag{7.12.3}$$

and to determine the numbers a_k by solving the system of non-linear equations

$$(y_n'' + f(x, y_n, y_n'), u_k) = 0\ ,\ \ k = 1, 2, \ldots, n\,. \tag{7.12.4}$$

If we can solve this system for each n, and if $y_n \in S_\delta(y_0)$ for all n sufficiently large, it can be shown that $y_n \to y_0$ as $n \to \infty$. Moreover, it can then be shown that

$$\|y_n - y_0\| = 0\left(\frac{1}{n^2}\,\|y_n'' + f(x, y_n, y_n')\|\right) \tag{7.12.5}$$

as $n \to \infty$.

Sometimes it is convenient to obtain accurate approximate solutions using a different choice of a sequence $\{u_k\}$, such as $u_k(x) = x(1-x)P_k(x)$, where the polynomials $P_k(x)$ are orthogonal over the interval $[0, 1]$ with respect to the weight function $x^2(1-x)^2$.

Example 7.12.1 Let us use the Bubnov-Galerkin method to obtain an approximate solution of

$$y'' + y + \frac{1}{2}\, y^3 + \frac{1}{8}\, \sin \pi x = 0 \ , \ \ y(0) = y(1) = 0 \tag{7.12.6}$$

having the form

$$y_1 = a \sin(\pi x)\,.$$

Setting

$$\int_0^1 \left[y_1'' + y_1 + \frac{1}{2}\, y_1^3 + \frac{1}{8}\, \sin \pi x \right] \sin \pi x dx = 0$$

we arrive at the equation

$$\frac{1}{2}\,(\pi^2 - 1)a - \frac{3}{16}\, a^3 - \frac{1}{16} = 0$$

that has the approximate solution $a \simeq \frac{1}{8(\pi^2-1)}$, so that $y_1 = \frac{\sin \pi x}{8(\pi^2-1)}$. It can be shown that this approximation is accurate to five decimals. Indeed

$$\left\| y_1'' + y_1 + \frac{1}{2}\, y_1^3 + \frac{1}{8}\, \sin \pi x \right\| < 3 \times 10^{-6}\,.$$

Further discussions on the use of the Bubnov-Galerkin method for solving ordinary differential equations can be found in [4, 18, 30, 31]. With proper choice of the basis functions u_k, the rate of convergence of y_n to y_0 is usually very rapid. The chief practical difficulty of the method is the solution of the system nonlinear equations (7.12.3).

Problems

7.12.1 Let the Hilbert space H be defined by the class of all functions f such that

$$\|f\|^2 = \int_0^\infty |f(x)|^2 dx < \infty\,.$$

Use the Bubnov-Galerkin method to find an approximate solution of the problem

$$y'' = y - \frac{y^3}{x^2} \ , \ \ y(0) = y(\infty) = 0$$

in the form

$$y_1 = a(e^{-x} - e^{-2x})\,.$$

7.12.2 Set $u_1 = 30^{1/2} \times (1 - x)$, $y_1 = au_1$ in (7.12.6) and determine the constant a such that (7.12.4) is satisfied with (u, v) defined in (7.12.2). How does the resulting solution compare with that obtained in the example of this section?

7.13 The Error of Numerical Methods for Solving Ordinary Differential Equations

Let us assume that the conditions of Theorem 7.1.10 are satisfied, namely, that $f(x, y)$ is continuous and satisfies a Lipschitz condition on a sufficiently large neighborhood of (x_0, y_0) that contains both the actual and the numerical solution of the problem

$$y' = f(x, y) \ , \ \ y(x_0) = y_0 \, . \tag{7.13.1}$$

We examine the error of Euler's method for this case [2].

Euler's method predicts an approximate solution y_p that satisfies the equation

$$y_{p+1} = y_p + hf(x_p, y_p) \quad , \quad p = 0, 1, 2, \ldots . \tag{7.13.2}$$

In practice we do not compute y_p, but rather z_p, where z_p satisfies the equation

$$z_{p+1} = z_p + hf(x_0, z_p) - R_p \tag{7.13.3}$$

where R_p is the roundoff error at the p'th step. The exact solution Y_p satisfies the equation

$$Y_{p+1} = Y_p + hf(x_p, Y_p) + T_p \tag{7.13.4}$$

where T_p is the truncation error of Euler's method, and is given by

$$T_p = \frac{h^2}{2} \, Y''(\zeta) \ , \ \ x_p < \zeta < x_{p+1} \, .$$

By taking the difference of (7.13.3) and (7.13.4), we get

$$Y_{p+1} - Y_{p+1} = Y_p - y_p + h[f(x_p, Y_p) - f(x_p, y_p)] + R_p + T_p \, . \tag{7.13.5}$$

Our assumed Lipschitz condition on f (see (7.1.31)) now tells us that

$$|f(x_p, y_p) - f(x_p, Y_p)| \leq M|y_p - Y_p| \tag{7.13.6}$$

where M does not depend on x and y in the region of consideration. Hence setting $e_p = y_p - Y_p$, we get the inequality

$$|e_{p+1}| \leq |e_p|(1 + Mh) + K \; ; \; e_0 = 0 \quad , \tag{7.13.7}$$

where we have assumed that $|T_p + R_p| \leq K$. Then

$$\begin{aligned} |e_1| &\leq K \\ |e_2| &\leq K[1 + (1 + Mh)] \\ &\vdots \\ |e_p| &\leq K[1 + (1 + Mh) + (1 + Mh)^2 + \cdots + (1 + Mh)^{p-1}] \\ &\leq \frac{K[(1 + Mh)^p - 1]}{Mh} \leq \frac{e^{M(x_p - x_0)} - 1}{Mh} \end{aligned} \tag{7.13.8}$$

since $1 + Mh \leq e^{Mh}$.

Notice that if we ignore roundoff error, then $K \leq Ch^2$, and $e_p = 0(h)$ as $h \to 0$. For moderate values of h, the roundoff error is considerably less than the truncation error. As h decreases, the total error decreases at first, and then becomes unbounded as $h \to 0$. A characteristic graph of the total error bound as a function of h is given in Figure 7.13.1.

We omit the analyses of more accurate methods, covered in the previous sections, since the analyses of these methods are similar to that of the Euler method.

Problems

7.13.1 For the case $y' = f(x, y)$, show that the local truncation error in the fourth order Runge-Kutta method is bounded by

$$MN(3.7 + 5.4M + 1.3M^2 + 0.017M^3)h^5$$

where $|f(x, y)| \leq N$, and

$$\left| \frac{f^{(i+j)}(x, y)}{\partial x^i \partial y^j} \right| \leq M/N \, .$$

7.13.2 Use the results of Problem 7.13.1 to carry out the error analysis of Section 6.13 for the case of the fourth order Runge-Kutta method of (7.6.10)–(7.6.11).

Figure 7.13.1: The Relationship Between Step Size and Error Bound

7.14 Stability of Linear Multistep Methods and Stiff Problems

The general k-step method for solving the problem

$$y' = f(x, y) \ , \ y(x_0) = y_0 \tag{7.14.1}$$

takes the form

$$y_{p+1} + \alpha_1 y_p + \cdots + \alpha_k y_{p-k+1} = h(\beta_0 f_{p+1} + \cdots + \beta_k f_{p-k+1}) \tag{7.14.2}$$

where $f_p = (f x_p, y_p)$, $k > 1$, the α_s and β_s are real constants, and $|\alpha_s| + |\beta_s| > 0$. In practice we do not solve the equation (7.14.2) for y_{p+1}, but rather the equation

$$Z_{p+1} + \alpha_1 Z_p + \cdots + \alpha_k Z_{p+1-k} = h(\beta_0 F_{p+1} + \cdots + \beta_k F_{p-k+1}) + R_p \tag{7.14.3}$$

where $F_p = f(x_p, Z_p)$, and where R_p denotes a roundoff error.

Roughly speaking, the equation (7.14.2) is called stable if the y_p differ only slightly from the Y_p, if all the R_p are small, where Y denotes the solution of (7.14.1). The general solution of the difference equation

$$z_{p+1} + \alpha_1 z_p + \cdots + \alpha_p z_{p+1-k} = 0 \tag{7.14.4}$$

takes the form

$$z_p = \sum_{i=1}^{q} \sum_{j=1}^{q_i-1} A_{ij} \zeta_i^p p^j \tag{7.14.5}$$

where the A_{ij} are constants, and where the ζ_1 are the $q \le k$ distinct roots of the polynomial equation

$$\zeta^k + \alpha_1 \zeta^{k-1} + \cdots + \alpha_k = 0\,. \tag{7.14.6}$$

If $R_p \not\equiv 0$, each term Z_p in (7.14.3) contains unwanted extraneous terms on the right-hand side of (7.14.5). If some of the ζ_1 are larger than 1 in modulus, or else if $|\zeta_1| = 1$ and $q_1 > 1$ for some i then this unwanted solution grows and eventually destroys the accuracy of the desired solution.

The above line of reasoning suggests the following definition of stability of the k-step method (7.14.2).

Definition 7.14.1 *The multistep method (7.14.2) is called stable if and only if all roots ζ_1 of (7.14.6) are at most 1 in modulus, and all roots of modulus 1 are simple.*

The above definition can be related more precisely to one involving the growth of error of the method (7.14.2). Let $f(x, y)$ be a continuous function of x and y for all $x \in [x_0, a]$ and $y \in \left\{y : \min_{x\in[x_0,a]} Y(x) - \alpha \le y \le \max_{[x_0,a]} Y(x) + \alpha\right\}$ where $Y(x)$ denotes the solution of (6.1), $\alpha > 0$, and let f satisfy the Lipschitz condition (1.28).

Definition 7.14.2 *We say that the perturbations R_p in (7.14.3) are of class $C(\delta)$ if, for some $h_0 > 0$,*

$$\sum_{p=0}^{k-1} |Z_p - y_p| + \sum_{j=k-1}^{N-1} |R_p| \le \delta \tag{7.14.7}$$

for all $h \le h_0$ and all N with $x_0 + Nh < a$.

For example, perturbations such that

$$|Z_p - y_p| \le \delta/2k\ ,\ |R_p| \le \delta h/[2(a - x_0)]$$

are of class $C(\delta)$.

Definition 7.14.3 *The multistep method (7.14.2) is called stable with respect to $f(x,y)$ if for any $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that the errors produced by perturbations of class $C(\delta)$ satisfy $|Z_p - y_p| \leq \epsilon (p = k, k+1, \ldots, N)$, uniformly for all $h \leq h_0$ and all N with $x_0 + Nh < a$.*

Theorem 7.14.4 *[39]: The multistep method (7.14.2) is stable in the sense of Definition 7.14.3 if and only if it is stable in the sense of Definition 7.14.1.*

We omit the proof of this theorem and refer the reader to [40, 6].

Definition 7.14.5 *The multistep method (7.14.2) is called consistent if*

$$1 + \alpha_1 + \alpha_2 + \cdots + \alpha_k = 0\,,$$

i.e., if $\zeta = 1$ is a solution of (7.14.6).

It may be shown [42] that every stable and consistent method of order greater than 1 is convergent.

We remark that by making the precision of computation sufficiently large the roundoff error of any k-step method can be made as small as we please, and hence the extraneous error can be made tolerably small. However, if the interval of integration is large, this usually demands an h so small and a precision so high that practical computation becomes unfeasible. Consider for example the following system of stiff equations:

$$\begin{bmatrix} f' \\ g' \end{bmatrix} = \begin{bmatrix} -11 & \frac{9}{12} \\ 18 & -11 \end{bmatrix} \begin{bmatrix} f \\ g \end{bmatrix}. \tag{7.14.8}$$

Such a system is called stiff because the eigenvalues θ_j of the coefficient matrix have negative real parts, such that the ratio of the largest modulus of the θ_j to the smallest modulus of the θ_j is large, e.g. greater than 10.

The eigenvalues of the coefficient matrix in (7.14.8) are -2 and -20. The solution with initial values (f_0, g_0) is

$$\begin{bmatrix} f(t) \\ g(t) \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\left(f_0 + \frac{1}{2}\,g_0\right) & \frac{1}{2}\left(f_0 - \frac{1}{2}\,g_0\right) \\ f_0 + \frac{1}{2}\,g_0 & -\left(f_0 - \frac{1}{2}\,g_0\right) \end{bmatrix} \begin{bmatrix} e^{-2t} \\ e^{-20t} \end{bmatrix}. \tag{7.14.9}$$

If we supply Euler's method to (7.14.8) we get the recurrence relation

$$\begin{bmatrix} f_{p+1} \\ g_{p+1} \end{bmatrix} = (I + Ah) \begin{bmatrix} f_p \\ g_p \end{bmatrix} \tag{7.14.10}$$

where I is the unit matrix of order 2, and A denotes the coefficient matrix of the system in (7.14.8). Thus

$$\begin{bmatrix} f_p \\ g_p \end{bmatrix} = (I+hA)^p \begin{bmatrix} f_0 \\ g_0 \end{bmatrix} = B^{-1} \begin{bmatrix} 1-2h & \\ & 1-20h \end{bmatrix}^p B \begin{bmatrix} f_0 \\ g_0 \end{bmatrix} \quad (7.14.11)$$

where

$$B = \begin{bmatrix} 1 & \frac{1}{2} \\ -1 & \frac{1}{2} \end{bmatrix} B^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ 1 & 1 \end{bmatrix} . \quad (7.14.12)$$

For example if $f_0 = 3$, $g_0 = 2$, then

$$f(t) = 2e^{-2t} + e^{-20t} \; , \; g(t) = 4e^{-2t} - 2e^{-20t} \quad (7.14.13)$$

and

$$f_p = 2(1-2h)^p + (1-20h)^p \; , \; g_p = 4(1-2h)^p - 2(1-20h)^p . \quad (7.14.14)$$

The e^{-20t} component decreases very rapidly relative to the e^{-2t} component, and hence in order to obtain the same behavior in the numerical solution, the step size h must be less than 1/20. However, after a few steps $(1-20h)^p$ is so small relative to $(1-2h)^p$ that it is comparable to the round-off error. At this stage we would like to be able to increase h, in order to be able to efficiently compute a solution over a larger interval. Note, however, that we cannot take $h > 0.1$, since then the $(1-20h)^p$ component would increase in magnitude. Hence we are forced to keep h small. Clearly this is an undesirable feature of the Euler method.

We are thus led to introduce the following notion of A-stability of a k-step method:

Definition 7.14.6 *The k-step method (7.14.2) is called A-stable when applied to (7.14.1), if every solution y_p of (7.14.2) converges to zero as $p \to \infty$ for all $h > 0$. The k-step method (7.14.2) is called explicit if $\beta_0 = 0$. It can be shown [32] that an explicit k-step method cannot be A-stable.*

For example, the Euler method is an explicit two step method, and hence is not A-stable. The trapezoidal method is A-stable. When applied to the problem

$$y' = -y$$

the trapezoidal method

$$y_{p+1} = y_p + \frac{h}{2}\left[f_{p+1} + f_p\right]$$

becomes

$$y_{p+1} = \left(\frac{1-h/2}{1+h/2}\right) y_p \, .$$

Clearly the modulus of $(1-h/2)/(1+h/2)$ is less than one for all $h>0$ and hence $y_p \to 0$ as $p \to \infty$.

It can be shown [22] that the order k of an A-stable consistent k-step method cannot be greater than 2.

Further Reading

Ordinary Differential Equations

J.F. Botha and G.F. Pinder. *Fundamental Concepts in the Numerical Solution of Differential Equations.* John Wiley and Sons, New York, 1987.

W.E. Boyce and R. C. DiPrima. *Elementary Differential Equations and Boundary Value Problems*, John Wiley and Sons, New York, 1965.

G.W. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971.

H.B. Keller. *Numerical Methods for Two-Point Boundary-Value Problems.* Blaisdell Publishing Co., Waltham, Massachusetts, ?year?.

E.L. Ince. *Ordinary Differential Equations.* Dover Publications, Inc., New York, 1956.

J.D. Lambert. *Computational Methods in Ordinary Differential Equations.* John Wiley and Sons, New York, 1973.

I. Stakgold. *Boundary Value Problems of Mathematical Physics*, volume I. MacMillan Co., New York, 1967.

I. Stakgold. *Boundary Value Problems of Mathematical Physics*, volume II. MacMillan Co., New York, 1967.

Partial Differential Equations

W.F. Ames. *Numerical Methods for Partial Differential Equations.* Academic Press, New York, 2nd edition, 1977.

D.A. Anderson, J.C. Tannenhill, and R.H. Pletcher. *Computational Fluid Mechanics and Heat Transfer.* McGraw-Hill Book Company, Wasington, 1984.

G. Birkhoff and R.E. Lynch. *Numerical Solution of Elliptic Problems.* SIAM, Philadelphia, 1984.

C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang. *Spectral Methods in Fluid Dynamics.* Springer-Verlag, New York, 1987.

D. Gottlieb and S.Z. Orszag. *Numerical Analysis of Spectral Methods: Theory and Applications.* SIAM, Philadelphia, 1977.

L. Lapidus and G.F. Pinder. *Numerical Solution of Partial Differential Equations in Science and Engineering.* John Wiley and Sons, New York, 1982.

R. Peyret and T.D. Taylor. *Computational Methods for Fluid Flow.* Springer-Verlag, New York, 1983.

G. D. Smith. *Numerical Solution of Partial Differential Equations: Finite Difference Methods.* Clarendon Press, Oxford, 2nd edition, 1978.

E.C. Zachmanoglou and D.W. Thoe. *Introduction to Partial Differential Equations.* Dover Publications, Inc., New York, 1986.