Numerical Computation I

E. Süli

Oxford University Computing Laboratory
Oxford 1998.

Preface. These lecture notes provide an introduction to computational methods for the approximation of functions on an interval of the real line. Only minimal prerequisites in differential and integral calculus, differential equation theory and linear algebra are necessary. The approach is aimed at giving an understanding of the construction of numerical algorithms and the analysis of their behaviour. The notes cover the material that appears in the first part of the course Numerical Computation, in Michaelmas Term. We begin with a study of methods and errors associated with the interpolation of functions by polynomials of given degree. We then use these techniques, in Section 2, for the derivation of numerical integration rules and their error analysis. Section 3 is devoted to the question of best approximation of a function by polynomials in the L^∞ and the L^2 norm; in particular, we shall describe the use of orthogonal polynomials for the construction of polynomials of best approximation in the L^2 norm, as well as their relevance in deriving Gauss-type numerical integration rules. In Section 4 we shall turn our attention to interpolation by piecewise polynomials of low degree – such objects are called splines. We conclude, in Section 5, with a brief overview of techniques for the numerical solution of initial value problems for ordinary differential equations.

Syllabus: - Numerical Computation I (12 lectures).
Interpolation of functions: Lagrange and Hermite interpolation, applications to quadrature, error analysis. Global polynomial approximation in the L^∞ and L^2 norm: inner product spaces; orthogonal polynomials, Gauss quadrature. Piecewise polynomial approximation: linear and Hermite cubic splines, B-splines.

Approximation of initial value problems for ordinary differential equations: one-step methods including Euler and Runge-Kutta methods; linear multi-step methods. Consistency, stability and convergence.

Reading List:

[1] Atkinson, K.E., *An Introduction to Numerical Analysis*, 2nd ed., Wiley 1989.

[2] Conte S.D. & de Boor, C., *Elementary Numerical Analysis*, 3rd ed, McGraw-Hill, 1980.

[3] Johnson, L.W. & Reiss, R.D., *Numerical Analysis*, Addison Wesley, 1977.

[4] Phillips, G.M. & Taylor, P.J., *Theory and Applications of Numerical Analysis, Academic Press, 1996.*

Further Reading:

[1] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962.

[2] K.W. Morton, *Numerical Solution of Ordinary Differential Equations*, Oxford University Computing Laboratory, 1987.

1 Interpolation of Functions

In this section we consider the problem of polynomial interpolation; this involves finding a polynomial that agrees exactly with some information that we have about the values of the function under consideration. The information may be in the form of values of the function at some set of points, and the corresponding polynomial is then called the Lagrange interpolation polynomial, or may include values of the derivatives of that function, in which case the associated polynomial is referred to as a Hermite interpolation polynomial.

1.1 Lagrange interpolation

Given that n is a non-negative integer, let $\mathcal{P}_n$ denote the set of polynomials of degree $\leq n$. The basic interpolation problem can be formulated as follows:

(**L**) Suppose that x_i, $i = 0, ..., n$, are *distinct* real numbers (i.e. $x_i \neq x_j$ for $i \neq j$), and let y_i, $i = 0, ..., n$, be real numbers; find $p_n \in \mathcal{P}_n$ such that $p_n(x_i) = y_i$, $i = 0, ..., n$.

We shall prove that problem (**L**) has a unique solution.

Uniqueness. Let us begin by showing that there is at most one polynomial p_n that satisfies the conditions formulated in (**L**). For suppose, otherwise, that there exists $q_n \in \mathcal{P}_n$, different from p_n, such that $q_n(x_i) = y_i$, $i = 0, ..., n$. Then $p_n - q_n \in \mathcal{P}_n$ and $p_n - q_n$ has $(n+1)$ distinct roots, x_i, $i = 0, ..., n$; therefore $p_n(x) - q_n(x) \equiv 0$, which contradicts our assumption that p_n and q_n are distinct. Consequently, there is at most one polynomial that solves problem (**L**).

Existence. Now we turn to showing the existence of a polynomial that satisfies the conditions in (**L**). The next lemma will be helpful.

Lemma 1 *Let $L_k \in \mathcal{P}_n$, $k = 0, ..., n$, be such that*

$$L_k(x_i) = \begin{cases} 1 & i = k \\ 0 & i \neq k. \end{cases}$$

Then

$$p_n(x) = \sum_{k=0}^{n} L_k(x) y_k \tag{1}$$

*satisfies the conditions formulated in (**L**), i.e. $p_n \in \mathcal{P}_n$ and $p_n(x_i) = y_i$, $i = 0, ..., n$.*

Remark 1 *By virtue of this lemma the existence of the polynomial p_n hinges on the existence of the L_k. In order to proceed we shall suppose for a moment that the polynomials L_k, satisfying the conditions of Lemma 1, exist: we shall prove later that this is indeed the case.*

PROOF (OF LEMMA 1): Since, by hypothesis, $L_k \in \mathcal{P}_n$, $k = 0, ..., n$, the linear combination of these polynomials, p_n, is also an element of $\mathcal{P}_n$; furthermore,

$$p_n(x_i) = \sum_{k=0}^{n} L_k(x_i) y_k = y_i, \tag{2}$$

and that completes the proof. □

As indicated in Remark 1, we still have to show that the polynomials L_k, $k = 0, ..., n$, exist. This is easily accomplished by explicitly constructing them. For each fixed k, $0 \leq k \leq n$, L_k is required to have n zeros, x_i, $i = 0, ..., n$, $i \neq k$; thus $L_k(x)$ is of the form

$$L_k(x) = C_k \prod_{i=0, i\neq k}^{n} (x - x_i), \tag{3}$$

where C_k is a constant. It is easy to determine the value of C_k by recalling that $L_k(x_k) = 1$; this yields

$$C_k = \prod_{i=0, i\neq k}^{n} (x_k - x_i)^{-1}.$$

Inserting this into (3) we obtain

$$L_k(x) = \prod_{i=0, i\neq k}^{n} \frac{x - x_i}{x_k - x_i}. \tag{4}$$

Thus, to summarise, the (unique) polynomial that solves problem (**L**) is given by (1) where the L_k, $k = 0, ..., n$, are defined by (4).

Definition 1 *Let n be a non-negative integer, let x_i, $i = 0, ..., n$, be distinct real numbers, and y_i, $i = 0, ..., n$, real numbers. The polynomial*

$$p_n(x) = \sum_{k=0}^{n} L_k(x) y_k,$$

with $L_k(x)$ defined by (4), is called the **Lagrange interpolation polynomial**[1] *of degree n for the set of points $\{(x_i, y_i) : i = 0, ..., n\}$. The numbers x_i, $i = 0, ..., n$, are called the* **interpolation points**.

Frequently the real numbers y_i are given as the values of a real-valued function f, defined on a closed real interval $[a, b]$, at the (distinct) interpolation points $x_i \in [a, b]$, $i = 0, ..., n$; in this case, $y_i = f(x_i)$, $i = 0, ..., n$, and the corresponding Lagrange interpolation polynomial has the form

$$p_n(x) = \sum_{k=0}^{n} L_k(x) f(x_k);$$

[1] Joseph-Louis Lagrange (1736–1813): Leçons élémentaires sur les mathématiques, Paris, 1795; Edward Warring (1734–1798) discovered the same interpolation formula in 1776.

the polynomial p_n is referred to as the Lagrange interpolation polynomial of degree n (with interpolation points x_i, $i = 0, ..., n$,) for the function f.

Although the values of the function f and those of its Lagrange interpolation polynomial coincide at the interpolation points, $f(x)$ may be quite different from $p_n(x)$ when x is not an interpolation point. Thus it is natural to ask just how large the difference $f(x) - p_n(x)$ is when $x \neq x_i$, $i = 0, ..., n$. Assuming that the function f is sufficiently smooth, an estimate of the size of the **interpolation error** $f(x) - p_n(x)$ is given in the next theorem.

Theorem 1 (A. L. Cauchy[2] (1840)) *Suppose that f is a real-valued function defined on the closed real interval $[a, b]$ and such that the derivative of f of order $(n + 1)$ is continuous on $[a, b]$. Suppose further that x_i, $i = 0, ..., n$, are distinct points in $[a, b]$. Then, given that $x \in [a, b]$,*

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\pi_{n+1}(x), \tag{5}$$

where $\xi = \xi(x) \in (a, b)$ and $\pi_{n+1}(x) = (x - x_0)...(x - x_n)$. Moreover,

$$|f(x) - p_n(x)| \leq \frac{M_{n+1}}{(n+1)!}|\pi_{n+1}(x)|, \tag{6}$$

where $M_{n+1} = \max_{\zeta \in [a,b]} |f^{(n+1)}(\zeta)|$.

PROOF: When $x = x_i$ for some i, $i = 0, ..., n$, both sides in (5) are equal to zero, and the equality is then trivially satisfied. Let us deal with the non-trivial case when $x \in [a, b]$ and $x \neq x_i$, $i = 0, ..., n$. For such x, let us consider the auxiliary function $t \mapsto \phi(t)$, defined on the interval $[a, b]$ by

$$\phi(t) = f(t) - p_n(t) - \frac{f(x) - p_n(x)}{\pi_{n+1}(x)}\pi_{n+1}(t).$$

Clearly, $\phi(x_i) = 0$, $i = 0, ..., n$, and $\phi(x) = 0$. Thus ϕ vanishes at $(n+2)$ points which are all distinct in $[a, b]$. Consequently, by Rolle's Theorem[3], $\phi'(t)$, the first derivative of ϕ with repect to t, vanishes at $(n + 1)$ points in (a, b), one point between each pair of consecutive points at which ϕ vanishes. Applying Rolle's Theorem again, we see that ϕ'' vanishes at n distinct points, and so on. Our assumptions about f are sufficient to apply Rolle's Theorem $(n + 1)$ times in succession, showing that $\phi^{(n+1)}$ vanishes at some point $\xi \in (a, b)$, the exact location of ξ being dependent on the position of x in (a, b). By differentiating $(n + 1)$ times the function ϕ with respect to t, and noting that p_n is a polynomial of degree n, it follows that

$$0 = \phi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{f(x) - p_n(x)}{\pi_{n+1}(x)}(n + 1)!.$$

[2] Augustin-Louis Cauchy (1789–1857)

[3] **Rolle's Theorem** (M. Rolle (1652–1719)): Suppose that the function f is defined and continuous on the closed real interval $[a, b]$, has a finite derivative $f'(x)$ at each point x in the open interval (a, b), and $f(a) = f(b)$. Then there exists at least one point ξ in (a, b) such that $f'(\xi) = 0$.

Thus,

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\pi_{n+1}(x).$$

In order to prove (6) let us note that since $f^{(n+1)}$ is a continuous function on $[a,b]$ the same is true of $|f^{(n+1)}|$; therefore $|f^{(n+1)}|$ is bounded on $[a,b]$ and achieves its maximum there. Denoting $M_{n+1} = \max_{\zeta\in[a,b]} |f^{(n+1)}(\zeta)|$, the inequality (6) follows from (5). $\square$

It is perhaps worth noting that, since the location of ξ in the interval $[a,b]$ is not known, (5) is of little practical value; on the other hand, given the function f, an upper bound on the maximum value of $f^{(n+1)}$ is, at least in principle, possible to obtain, and thereby we can provide an upper bound on the size of the interpolation error by means of the inequality (6).

Exercise 1 *Let $f(x) = \mathrm{e}^x$ for $x \in [-1,1]$. Write down the Lagrange interpolation polynomial $p_2(x)$ of degree 2 with interpolation points -1, 0, 1 for the function f. Show further that*

$$|f(x) - p_2(x)| \le \frac{\mathrm{e}}{6}|x|(1-x^2)$$

for all x in $[-1,1]$.

SOLUTION: Letting $x_0 = -1$, $x_1 = 0$, $x_2 = 1$, the Lagrange interpolation polynomial of degree $n = 2$ for the function f is

$$p_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}f(x_2).$$

Thus,

$$p_2(x) = \frac{1}{2\mathrm{e}}x(x-1) + (1-x^2) + \frac{\mathrm{e}}{2}x(x+1),$$

or, upon rearrangement,

$$p_2(x) = x^2(\cosh 1 - 1) + x\sinh 1 + 1.$$

Now to prove the desired error bound, let us note that

$$M_3 = \max_{\xi\in[-1,1]} |f'''(\xi)| = \mathrm{e}$$

and $\pi_3(x) = (x-x_0)(x-x_1)(x-x_2) = x(x^2-1)$. Thus, from (6) we have that

$$|f(x) - p_2(x)| \le \frac{\mathrm{e}}{6}|x|(1-x^2),$$

as required. $\diamond$

Exercise 2 (Oxford Finals, 1992) *For a non-negative n, and x in $[-1,1]$, let $T_n(x) = \cos(n\cos^{-1}x)$. Deduce the recurrence relation*

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0, \quad n \geq 1.$$

Hence show that T_n is a polynomial of degree n and that, for $n \geq 1$, the leading term of $T_n(x)$ is $2^{n-1}x^n$. Let $f : [-1,1] \to \mathbf{R}$ be a continuous function and let x_i, $i = 0,\ldots,n$, denote the zeros of $T_{n+1}(x)$. Prove that there exists a unique polynomial p_n of degree n such that $p_n(x_i) = f(x_i)$, $i = 0,\ldots,n$.

Show that if $f^{(n+1)}$ exists and is a continuous function on the interval $[-1,1]$ then

$$\max_{x\in[-1,1]} |f(x) - p_n(x)| \leq \frac{M_{n+1}}{2^n(n+1)!},$$

where $M_{n+1} = \max_{\xi\in[-1,1]} |f^{(n+1)}(\xi)|$.

SOLUTION: Let $\phi = \cos^{-1}x$; then, for $n \geq 1$,

$$\begin{aligned} T_{n+1}(x) + T_{n-1}(x) &= \cos(n+1)\phi + \cos(n-1)\phi \\ &= 2\cos n\phi \cdot \cos\phi = 2x\cos(n\cos^{-1}x) = 2xT_n(x), \end{aligned}$$

which is the required recurrence. Writing

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),$$

and noting that $T_0(x) \equiv 1$, $T_1(x) = x$, it follows by induction that T_n is a polynomial of degree n. Indeed, this is true for $n = 0$ and $n = 1$; let us suppose that T_n is a polynomial of degree n for all n, $n \leq k$; then, by the recurrence relation, T_{k+1} is a polynomial and

$$\text{degree of } T_{k+1} = \text{degree of } (2xT_k - T_{k-1}) = k+1.$$

That completes the induction.

Also, by induction, the leading term of $T_n(x)$ is $2^{n-1}x^n$. Indeed, this is true for $n = 1$ as $x = 2^{1-1}x$; let us suppose that the leading term of $T_k(x)$ is $2^{k-1}x^k$ for some k, $k \geq 1$; then

$$\begin{aligned} &\text{leading term of } T_{k+1}(x) = \text{leading term of } 2xT_k(x) \\ &\quad = 2x(\text{leading term of } T_k(x)) = 2^k x^{k+1}, \end{aligned}$$

and that completes the induction.

Let x_i, $i = 0,\ldots,n$, denote the zeros of $T_{n+1}(x)$:

$$x_i = \cos\frac{(2i+1)\pi}{2n+2}, \quad i = 0,\ldots,n.$$

These are distinct real numbers in the closed interval $[-1,1]$ because

$$\frac{(2i+1)\pi}{2n+2} \in [0,\pi], \quad i = 0,\ldots,n,$$

and cos is strictly monotonic decreasing on the interval $[0, \pi]$.

Now letting $y_i = f(x_i)$, $i = 0, \ldots, n$, and repeating the argument from the beginning of the section it follows that there exists a unique polynomial p_n of degree n such that $p_n(x_i) = y_i = f(x_i)$. Further, from (6) it follows that

$$\max_{x\in[-1,1]} |f(x) - p_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \max_{x\in[-1,1]} |\pi_{n+1}(x)|.$$

As

$$T_{n+1}(x) = 2^n(x - x_0)\ldots(x - x_n) = 2^n \pi_{n+1}(x),$$

and $\max_{x\in[-1,1]} |T_{n+1}(x)| = 1$, we have that

$$\max_{x\in[-1,1]} |\pi_{n+1}(x)| = 2^{-n} \max_{x\in[-1,1]} |T_{n+1}(x)| = 2^{-n},$$

and hence the required inequality. $\diamond$

We note here that $T_n(x)$ is called a Chebyshev polynomial of degree n. Chebyshev polynomials play an important rôle in approximation theory; we shall return to the study of their properties later on in these notes when we discuss the problem of best approximation.

1.2 Hermite interpolation

The idea of Lagrange interpolation can be generalised in various ways; we shall consider here one simple extension where a polynomial p is required to take given values and derivative values at the interpolation points. Let us formalise this interpolation problem.

(**H**) Suppose that n is a non-negative integer and let x_i, $i = 0, ..., n$, be distinct real numbers. Given two sets of real numbers y_i, $i = 0, ..., n$, and z_i, $i = 0, ..., n$, find a polynomial $p_{2n+1} \in \mathcal{P}_{2n+1}$ such that

$$p_{2n+1}(x_i) = y_i, \qquad p'_{2n+1}(x_i) = z_i, \qquad i = 0, ..., n.$$

As in the case of Lagrange interpolation, we begin by showing that problem (**H**) has a unique solution.

Uniqueness. First we shall prove that there is at most one polynomial that satisfies the conditions formulated in (**H**). Suppose otherwise: then there exists a polynomial $q_{2n+1} \in \mathcal{P}_{2n+1}$, distinct from p_{2n+1}, such that $q_{2n+1}(x_i) = y_i$ and $q'_{2n+1}(x_i) = z_i$, $i = 0, ..., n$. Consequently, $p_{2n+1} - q_{2n+1}$ has $(n + 1)$ distinct zeros; thus Rolle's Theorem implies that, in addition to the $(n + 1)$ zeros x_i, $i = 0, ..., n$, $p'_{2n+1} - q'_{2n+1}$ vanishes at another n points which interlace the x_i. Hence $p'_{2n+1} - q'_{2n+1} \in \mathcal{P}_{2n}$ has $(2n+1)$ zeros, which means that $p'_{2n+1} - q'_{2n+1}$ is identically zero, so that $p_{2n+1} - q_{2n+1}$ is a constant function; but $p_{2n+1} - q_{2n+1}$ vanishes at x_i, $i = 0, ..., n$, and therefore $p_{2n+1} - q_{2n+1} \equiv 0$, contradicting to the hypothesis that p_{2n+1} and q_{2n+1} are distinct. Thus, if it exists, p_{2n+1} is unique.

Existence. Let us consider the polynomials H_k and K_k, $k = 0, ..., n$, defined by

$$H_k(x) = [L_k(x)]^2 \left(1 - 2L_k'(x_k)(x - x_k)\right), \tag{7}$$
$$K_k(x) = [L_k(x)]^2 (x - x_k), \tag{8}$$

where

$$L_k(x) = \prod_{i=0, i\neq k}^{n} \frac{x - x_i}{x_k - x_i}.$$

Clearly H_k and K_k, $k = 0, ..., n$, are polynomials of degree $(2n + 1)$; moreover, a straightforward calculation shows that

$$H_k(x_i) = \begin{cases} 1 & i = k \\ 0 & i \neq k, \end{cases} \qquad H_k'(x_i) = 0, \qquad i, k = 0, ..., n,$$

$$K_k(x_i) = 0, \qquad K_k'(x_i) = \begin{cases} 1 & i = k \\ 0 & i \neq k, \end{cases} \qquad i, k = 0, ..., n.$$

Thus, the polynomial

$$p_{2n+1}(x) = \sum_{k=0}^{n} H_k(x) y_k + K_k(x) z_k$$

satisfies the conditions formulated in (**H**); furthermore, by what has been proved above, it is the unique such polynomial.

Definition 2 *Let n be a non-negative integer, let x_i, $i = 0, ..., n$, be distinct real numbers, and y_i, z_i, $i = 0, ..., n$, real numbers. The polynomial*

$$p_{2n+1}(x) = \sum_{k=0}^{n} H_k(x) y_k + K_k(x) z_k,$$

where $H_k(x)$ and $K_k(x)$ are defined by (7), (8) respectively, is called the **Hermite interpolation polynomial**[4] *of degree $(2n + 1)$ for the set of points $\{(x_i, y_i) : i = 0, ..., n\}$, $\{(x_i, z_i) : i = 0, ..., n\}$.*

Exercise 3 *Find a cubic polynomial p_3 such that $p_3(0) = 0$, $p_3(1) = 1$, $p_3'(0) = 1$, $p_3'(1) = 0$.*

SOLUTION: Letting $n = 1$, $x_0 = 0$, $x_1 = 1$, $y_0 = 0$, $y_1 = 1$, $z_0 = 1$, $z_1 = 0$, we seek $p_3(x)$, following Definition 2, as

$$p_3(x) = 0 \times H_0(x) + 1 \times H_1(x) + 1 \times K_0(x) + 0 \times K_1(x).$$

Thus

$$p_3(x) = H_1(x) + K_0(x).$$

[4] Charles Hermite (1822-1901)

Now, with $x_0 = 0$ and $x_1 = 1$, we have that

$$\begin{aligned} H_1(x) &= [L_1(x)]^2(1 - 2L_1'(x_1)(x - x_1)), \\ K_0(x) &= [L_0(x)]^2(x - x_0). \end{aligned}$$

It remains to determine $L_0(x)$ and $L_1(x)$. Clearly,

$$L_0(x) = \frac{x - x_1}{x_0 - x_1} = \frac{x - 1}{0 - 1} = 1 - x,$$
$$L_1(x) = \frac{x - x_0}{x_1 - x_0} = \frac{x - 0}{1 - 0} = x.$$

Therefore,

$$H_1(x) = x^2(3 - 2x) \quad \text{and} \quad K_0(x) = (1 - x)^2 x,$$

so that

$$p_3(x) = -x^3 + x^2 + x$$

is the required Hermite interpolation polynomial. $\diamond$

Suppose that f is a real-valued function that is defined on the closed interval $[a, b]$ and is continuous and differentiable on this interval; suppose, further, that x_i, $i = 0, ..., n$, are distinct points in this interval. We can use Hermite interpolation to construct a polynomial $p_{2n+1} \in \mathcal{P}_{2n+1}$ which takes the same function values and derivative values as f. To do so it suffices to choose $y_i = f(x_i)$ and $z_i = f'(x_i)$, $i = 0, ..., n$, which yields

$$p_{2n+1}(x) = \sum_{k=0}^{n} H_k(x) f(x_k) + K_k(x) f'(x_k);$$

the polynomial p_{2n+1} is referred to as the Hermite interpolation polynomial of degree $2n + 1$ (with interpolation points x_i, $i = 0, ..., n,$) for f. Pictorially, the graph of p_{2n+1} touches the graph of the function f at the points x_i, $i = 0, ..., n$.

To conclude this section we state a result, analogous to Theorem 1, concerning the error in Hermite interpolation.

Theorem 2 *Suppose that f is a real-valued function defined on the closed interval $[a, b]$ and such that $f^{(2n+2)}$ is continuous on $[a, b]$. Suppose further that x_i, $i = 0, ..., n$, are distinct points in $[a, b]$. Then, given that $x \in [a, b]$,*

$$f(x) - p_{2n+1}(x) = \frac{f^{(2n+2)}(\xi)}{(2n + 2)!}[\pi_{n+1}(x)]^2, \tag{9}$$

where $\xi = \xi(x) \in (a, b)$ and $\pi_{n+1}(x) = (x - x_0)...(x - x_n)$. Moreover,

$$|f(x) - p_{2n+1}(x)| \leq \frac{M_{2n+2}}{(2n + 2)!}[\pi_{n+1}(x)]^2, \tag{10}$$

where $M_{2n+2} = \max_{\zeta \in [a,b]} |f^{(2n+2)}(\zeta)|$.

PROOF: The inequality (10) is a straightforward consequence of (9). In order to prove (9) let us observe that it is trivially true if $x = x_i$ for some i, $i = 0, ..., n$; thus it suffices to consider $x \in [a, b]$ such that $x \neq x_i$, $i = 0, ..., n$. For such x, let us define the function $t \mapsto \psi(t)$ by

$$\psi(t) = f(t) - p_{2n+1}(t) - \frac{f(x) - p_{2n+1}(x)}{[\pi_{n+1}(x)]^2}[\pi_{n+1}(t)]^2.$$

Then $\psi(x_i) = 0$ for $i = 0, ..., n$, and also $\psi(x) = 0$. Hence, by Rolle's Theorem, $\psi'(t)$ vanishes at $(n + 1)$ points which lie strictly between each pair of consecutive points from the set $\{x_0, ..., x_n, x\}$. Also, $\psi'(x_i) = 0$, $i = 0, ..., n$; hence ψ' vanishes at a total of $(2n + 2)$ distinct points in $[a, b]$. Applying Rolle's Theorem in succession, we find eventually that $\psi^{(2n+2)}$ vanishes at some point ξ in (a, b), the location of ξ being dependent on the position of x. This gives the required result since $p_{2n+1}^{(2n+2)}(x) \equiv 0$. □

2 Numerical Integration - Part I

In this section we apply the results of Section 1.1 to derive formulae for numerical integration (also called numerical quadrature rules) and estimate the associated approximation error.

2.1 Newton-Cotes formulae

Let f be a real-valued function, defined and continuous on a closed real interval $[a, b]$, and suppose that we have to evaluate the integral

$$\int_a^b f(x)\,\mathrm{d}x.$$

Since polynomials are easy to integrate, the idea, roughly speaking, is to approximate the function f by its Lagrange interpolation polynomial p_n of degree n, and integrate p_n instead. Thus,

$$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b p_n(x)\,\mathrm{d}x. \tag{11}$$

For a positive integer n, let x_i, $i = 0, ..., n$, denote the interpolation points; for the sake of simplicity, we shall assume that these are equally spaced, namely, $x_i = a + i\,h$ for $i = 0, ..., n$, where $h = (b - a)/n$. The Lagrange interpolation polynomial of degree n for the function f, with these interpolation points, is of the form

$$p_n(x) = \sum_{k=0}^{n} L_k(x) f(x_k).$$

Inserting this into the right-hand side of (11) yields

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{k=0}^{n} w_k f(x_k), \tag{12}$$

where

$$w_k = \int_a^b L_k(x)\,\mathrm{d}x, \qquad k = 0, ..., n,$$

are referred to as the **quadrature weights**, while the interpolation points x_k, $k = 0, ..., n$, are called the **quadrature points**. Numerical quadrature rules of this form, with equally spaced quadrature points, are called **Newton**[5] **– Cotes formulae**. In order to illustrate the general idea, we consider two simple examples.

Trapezium rule. In this case we take $n = 1$, $x_0 = a$ and $x_1 = b$; the Lagrange interpolation polynomial of degree 1 for the function f is simply

$$\begin{aligned} p_1(x) &= L_0(x)f(x_0) + L_1(x)f(x_1) \\ &= \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1) \\ &= \frac{1}{b-a}\left[(b-x)f(a) + (x-a)f(b)\right]. \end{aligned}$$

Integrating $p_1(x)$ from a to b yields

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2}\left[f(a) + f(b)\right].$$

This numerical integration formula is called the trapezium rule.

Simpson's rule[6]**.** A slightly more sophisticated quadrature rule is obtained by taking $n = 2$. In this case $x_0 = a$, $x_1 = (a+b)/2$ and $x_2 = b$; the Lagrange polynomial of degree 2, with these interpolation points, for the function f is:

$$\begin{aligned} p_2(x) &= L_0(x)f(x_0) + L_1(x)f(x_1) + L_2(x)f(x_2) \\ &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) \\ &\qquad + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2). \end{aligned}$$

[5] Isaac Newton (1643–1727)

[6] A straighforward calculation shows that the area under the arc of a parabola $f(x) = Ax^2 + Bx + C$ in an interval $[a, b]$ is given by the Cavalieri-Gregory formula:

$$\text{Area} = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

This formula was stated in geometric terms by B. Cavalieri (Centuria di varii problemi, Bologna 1639); it was independently discovered by J. Gregory (Exercitationes geometricae, London 1668), and Th. Simpson (Mathematical dissertations on a variety of physical and analytical subjects, London 1743). Surprisingly, the Cavalieri-Gregory formula is valid even for a cubic polynomial $f(x) = Ax^3 + Bx^2 + Cx + D$ – the explanation is supplied by Lemma 2 below.

Integrating $p_2(x)$ from a to b gives

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

This numerical integration formula is known as Simpson's rule.

Our next task is to estimate the size of the error in the numerical integration formula (12), that is, the error that has been committed by integrating the Lagrange interpolation polynomial of f instead of the function f itself. The error in (12) is defined by

$$E_n(f) = \int_a^b f(x)\,\mathrm{d}x - \sum_{k=0}^{n} w_k f(x_k).$$

The next theorem provides a useful bound on $E_n(f)$ under the additional hypothesis that the function f is sufficiently smooth.

Theorem 3 *Suppose that f is a real-valued function defined on the interval $[a,b]$, and let $f^{(n+1)}$ be continuous on $[a,b]$. Then,*

$$|E_n(f)| \leq \frac{M_{n+1}}{(n+1)!}\int_a^b |\pi_{n+1}(x)|\,\mathrm{d}x, \tag{13}$$

where $M_{n+1} = \max_{\zeta\in[a,b]} |f^{(n+1)}(\zeta)|$ and $\pi_{n+1}(x) = (x-x_0)...(x-x_n)$.

Proof: Recalling the definition of the weights w_k in (12), we can rewrite $E_n(f)$ as follows:

$$\begin{aligned} E_n(f) &= \int_a^b f(x)\,\mathrm{d}x - \int_a^b \left(\sum_{k=0}^{n} L_k(x) f(x_k)\right)\,\mathrm{d}x \\ &= \int_a^b (f(x) - p_n(x))\,\mathrm{d}x. \end{aligned}$$

Thus,

$$|E_n(f)| \leq \int_a^b |f(x) - p_n(x)|\,\mathrm{d}x.$$

The desired error estimate (13) follows by inserting (6) into the right-hand side of this inequality. □

Let us apply Theorem 3 to estimate the size of the error that has been committed by applying the trapezium rule to the integral $\int_a^b f(x)\,\mathrm{d}x$. In this case (13) reduces to

$$\begin{aligned} |E_1(f)| &\leq \frac{1}{2}M_2\int_a^b |(x-a)(x-b)|\,\mathrm{d}x \\ &= \frac{1}{2}M_2\int_a^b (b-x)(x-a)\,\mathrm{d}x \\ &= \frac{1}{12}(b-a)^3 M_2. \end{aligned} \tag{14}$$

An analogous but slightly more tedious calculation shows that, for Simpson's rule,

$$\begin{aligned} |E_2(f)| &\leq \frac{1}{6} M_3 \int_a^b |(x-a)(x-\frac{a+b}{2})(x-b)| \, \mathrm{d}x \\ &= \frac{1}{192} M_3 (b-a)^4. \end{aligned} \tag{15}$$

Unfortunately, (15) gives a very pessimistic estimate of the error in Simpson's rule; indeed, it is easy to verify that, whenever f is a polynomial of degree 3 we have $E_2(f) = 0$ while the right-hand side of (15) is a positive real number. The next lemma will allow us to sharpen this crude bound.

Lemma 2 *Suppose that f is a real-valued function defined on the interval $[a, b]$ and $f^{(4)}$ is a continuous function on $[a, b]$. Then*

$$\int_a^b f(x)\, dx - \frac{b-a}{6}\left[f(a) + f\left(\frac{a+b}{2}\right) + f(b)\right] = -\frac{(b-a)^5}{2880} f^{(4)}(\xi), \tag{16}$$

for some ξ in (a, b).

Proof: Performing the change of variables

$$x = \frac{a+b}{2} + \frac{b-a}{2} t, \qquad t \in [-1, 1],$$

and defining the function $t \mapsto F(t)$ by $F(t) = f(x)$, we have that

$$\begin{aligned} &\int_a^b f(x)\, \mathrm{d}x - \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \\ &\qquad = \frac{b-a}{2}\left(\int_{-1}^1 F(t)\, \mathrm{d}t - \frac{1}{3}\left[F(-1) + 4F(0) + F(1)\right]\right). \end{aligned} \tag{17}$$

Let us introduce

$$G(t) = \int_{-t}^t F(\tau)\, \mathrm{d}\tau - \frac{t}{3}[F(-t) + 4F(0) + F(t)], \qquad t \in [-1, 1];$$

then the right-hand side of (17) is simply $\frac{1}{2}(b-a)G(1)$.

The remainder of the proof is devoted to showing that $\frac{1}{2}(b-a)G(1)$ is, in turn, equal to the right-hand side of (16) for some ξ in (a, b). For this purpose, we define

$$H(t) = G(t) - t^5 G(1), \qquad t \in [-1, 1].$$

We shall apply Rolle's Theorem four times to the function H. Noting that $H(0) = H(1) = 0$, we deduce that there exists $\zeta_1 \in (0, 1)$ such that $H'(\zeta_1) = 0$. However, also, $H'(0) = 0$, so there exists $\zeta_2 \in (0, \zeta_1)$ such that $H''(\zeta_2) = 0$; but, $H''(0) = 0$,

so there is $\zeta_3 \in (0, \zeta_2)$ such that $H'''(\zeta_3) = 0$. Finally, noting that $H'''(0) = 0$, it follows that there is $\zeta_4 \in (0, \zeta_3)$ such that $H^{(4)}(\zeta_4) = 0$. Thus,

$$H^{(4)}(\zeta_4) = G^{(4)}(\zeta_4) - 120\zeta_4 G(1) = 0.$$

Now

$$G^{(4)}(t) = -\frac{1}{3}(F'''(t) - F'''(-t)) - \frac{t}{3}(F^{(4)}(t) + F^{(4)}(-t)),$$

and therefore

$$G(1) = -\frac{1}{360\zeta_4}(F'''(\zeta_4) - F'''(-\zeta_4)) - \frac{1}{360}(F^{(4)}(\zeta_4) + F^{(4)}(-\zeta_4)).$$

Applying Lagrange's Mean Value Theorem[7] to the first term on the right, we deduce that

$$G(1) = -\frac{1}{180}F^{(4)}(\zeta_5) - \frac{1}{360}(F^{(4)}(\zeta_4) + F^{(4)}(-\zeta_4)),$$

for some $\zeta_5 \in (-\zeta_4, \zeta_4)$. Equivalently,

$$G(1) = -\frac{1}{90}\left(\frac{F^{(4)}(-\zeta_4) + 2F^{(4)}(\zeta_5) + F^{(4)}(\zeta_4)}{4}\right).$$

Since, by our hypothesis on f, $F^{(4)}$ is a continuous function on $[-1, 1]$ and $-1 < -\zeta_4 < \zeta_5 < \zeta_4 < 1$, it follows that there exists θ in $[-\zeta_4, \zeta_4]$ such that

$$\frac{F^{(4)}(-\zeta_4) + 2F^{(4)}(\zeta_5) + F^{(4)}(\zeta_4)}{4} = F^{(4)}(\theta);$$

consequently,

$$G(1) = -\frac{1}{90}F^{(4)}(\theta) = -\frac{1}{1440}(b-a)^4 f^{(4)}(\xi), \tag{18}$$

where $\xi = \frac{1}{2}(a+b) + \frac{1}{2}(b-a)\theta$. Finally, (16) is obtained by inserting (18) into (17). □

Now Lemma 2 yields the following bound on the error in Simpson's rule:

$$|E_2(f)| \leq \frac{1}{2880}(b-a)^5 M_4. \tag{19}$$

This is a considerable improvement over (15) in the sense that if $f \in \mathcal{P}_3$ then the right-hand side of (19) is equal to zero and thereby $E_2(f) = 0$ which now correctly reflects the fact that polynomials of degree three are integrated by Simpson's rule

[7] **Mean Value Theorem:** Suppose that the function f is defined and continuous on the closed real interval $[a, b]$ and that f has finite derivative $f'(x)$ at each point x of the open interval (a, b). Then, there exists ζ in (a, b) such that

$$f(b) - f(a) = f'(\zeta)(b-a).$$

without error. (As remarked earlier, this property was not borne out by our initial crude bound (15).)

By considering the right-hand side of the error bound in Theorem 3 we may be led to believe that by increasing n, that is by approximating the integrand by Lagrange polynomials of increasing degree and integrating these exactly, we shall be reducing the size of the quadrature error $E_n(f)$. It is easy to find examples which show that this is not always the case[8]. A better approach is to divide the interval $[a, b]$ into an increasing number of subintervals (of decreasing size) and use a numerical integration formula of fixed order on each of the subintervals. Quadrature rules based on this approach are called composite formulae, and will be described in detail in the next section.

2.2 Composite formulae

We shall consider only some very simple composite quadrature rules: the composite trapezium rule, and the composite Simpson rule.

Composite trapezium rule. This is obtained by dividing the interval $[a, b]$ into m equal subintervals, each of width $h = (b - a)/m$, so that

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{m} \int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x,$$

where

$$x_i = a + ih = a + \frac{i}{m}(b - a), \qquad i = 0, ..., m.$$

Each of the integrals is then evaluated by the trapezium rule, namely,

$$\int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x \approx \frac{1}{2}h[f(x_{i-1}) + f(x_i)],$$

giving the complete approximation

$$\int_a^b f(x)\,\mathrm{d}x \approx h[\frac{1}{2}f(x_0) + f(x_1) + ... + f(x_{m-1}) + \frac{1}{2}f(x_m)], \tag{20}$$

called the **composite trapezium rule**.

The error in the composite trapezium rule can be estimated by using the error bound (14) for the trapezium rule on each individual subinterval $[x_{i-1}, x_i]$, $i = 1, ..., m$. Indeed,

$$\begin{aligned} \mathcal{E}_1(f) &:= \int_a^b f(x)\,\mathrm{d}x - h[\frac{1}{2}f(x_0) + f(x_1) + ... + f(x_{m-1}) + \frac{1}{2}f(x_m)] \\ &= \sum_{i=1}^{m} \left[\int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x - \frac{1}{2}h[f(x_{i-1}) + f(x_i)] \right]. \end{aligned}$$

[8] Consider, for example, $\int_{-5}^{5} 1/(1+x^2)\,\mathrm{d}x$, and use Newton-Cotes formulae for higher and higher values of n on the interval $[-5, 5]$. This pathological example was discovered by the German mathematician Karl Runge.

Applying (14) to each of the terms under the summation sign,

$$\begin{aligned} |\mathcal{E}_1(f)| &\leq \frac{1}{12}h^3 \sum_{i=1}^{m} \left(\max_{\zeta\in[x_{i-1},x_i]} |f''(\zeta)| \right) \\ &\leq \frac{1}{12m^2}(b-a)^3 M_2, \end{aligned} \tag{21}$$

where $M_2 = \max_{\zeta\in[a,b]} |f''(\zeta)|$.

Composite Simpson's rule. Let us suppose that the interval $[a,b]$ has been divided into $2m$ subintervals by the points $x_i = a + i\,h$, $i = 0, ..., 2m$, where $h = (b-a)/(2m)$, and let us apply Simpson's rule on each of the intervals $[x_{2i-2}, x_{2i}]$, $i = 1, ..., m$, giving

$$\begin{aligned} \int_a^b f(x)\,\mathrm{d}x &= \sum_{i=1}^{m} \int_{x_{2i-2}}^{x_{2i}} f(x)\,\mathrm{d}x \\ &\approx \sum_{i=1}^{m} \frac{2h}{6}\left[f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})\right]. \end{aligned}$$

Equivalently,

$$\begin{aligned} \int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{3}\big[&f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + ... \\ &+ 2f(x_{2m-2}) + 4f(x_{2m-1}) + f(x_{2m})\big]. \end{aligned} \tag{22}$$

The numerical integration formula (22) is called the **composite Simpson rule**.

In order to estimate the error in the composite Simpson rule, we proceed in the same way as for the composite trapezium rule:

$$\begin{aligned} \mathcal{E}_2(f) &:= \int_a^b f(x)\,\mathrm{d}x - \sum_{i=1}^{m} \frac{h}{3}\left[f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})\right] \\ &= \sum_{i=1}^{m} \left[\int_{x_{2i-2}}^{x_{2i}} f(x)\,\mathrm{d}x - \frac{h}{3}\left[f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})\right] \right]. \end{aligned}$$

Applying (19) to each individual term under the summation sign and recalling that $b - a = 2mh$, we obtain the following error bound:

$$|\mathcal{E}_2(f)| \leq \frac{1}{2880m^4}(b-a)^5 M_4. \tag{23}$$

The composite rules (20) and (22) provide greater accuracy than the basic formulae considered in Section 2.1; this is clearly seen by comparing the error bounds (21) and (23) for the two composite rules with (14) and (19), the error estimates for the basic trapezium– and Simpson formula, respectively. The inequalities (21) and (23) indicate that, as long as the function f is sufficiently smooth, the errors in the composite rules can be made arbitrarily small by choosing a sufficiently large number of subintervals.

Exercise 4 *Calculate the integral*

$$I = \int_0^{\pi/2} \sin x \, \mathrm{d}x$$

with accuracy $\epsilon = 10^{-2}$ *using:*

a) the composite trapezium rule;

b) the composite Simpson rule.

SOLUTION: a) In order to approximate the integral I to within the given accuracy ϵ by means of the composite trapezium rule, we have to find a positive integer m (the number of subdivisions of the interval $[0, \pi/2]$), as small as possible, such that

$$|\mathcal{E}_1(f)| \leq \epsilon,$$

where $f(x) = \sin x$ with $x \in [0, \pi/2]$ and $\epsilon = 10^{-2}$. Recalling the error bound for the composite trapezium rule, this amounts to finding the smallest positive integer m such that

$$\frac{1}{12m^2}(b-a)^3 M_2 \leq \epsilon,$$

where $M_2 = \max_{x\in[0,\pi/2]} |f''(x)| = 1$, $b - a = \pi/2 - 0 = \pi/2$. Thus we need to choose m, as small as possible, such that

$$\frac{1}{12m^2}\left(\frac{\pi}{2}\right)^3 \leq 10^{-2}.$$

The smallest integer m for which this inequality holds is $m = 6$. Having determined the required number of subdivisions, we calculate the approximation to I by the composite trapezium rule:

$$\begin{aligned} I &\approx \frac{\pi}{12}\left[\frac{1}{2}\sin 0 + \left(\sin\frac{\pi}{12} + \sin\frac{2\pi}{12} + \sin\frac{3\pi}{12} + \sin\frac{4\pi}{12} + \sin\frac{5\pi}{12}\right) + \frac{1}{2}\sin\frac{\pi}{2}\right] \\ &= 0.99429\ . \end{aligned}$$

b) For the composite Simpson rule we proceed similarly as in part a): we require that $|\mathcal{E}_2(f)| \leq \epsilon$, with as small a number of subdivisions as possible; namely, we want to find a positive integer m, as small as possible, such that

$$\frac{1}{2880m^4}(b-a)^5 M_4 \leq \epsilon,$$

where $b - a = \pi/2$, $M_4 = 1$ and $\epsilon = 10^{-2}$. Thus we demand that

$$\frac{1}{2880m^4}\left(\frac{\pi}{2}\right)^5 \leq 10^{-2}.$$

The smallest positive integer m for which this holds is $m = 1$. Having found m, we can calculate the approximation to I by means of the composite Simpson rule (which in this case, with $m = 1$, is simply the basic Simpson rule):

$$I \approx \frac{\pi}{12}\left[\sin 0 + 4\sin\frac{\pi}{4} + \sin\frac{\pi}{2}\right] = 1.00228.$$

We note that the exact value of the integral is $I = 1$. Hence both approximations are accurate to within the required accuracy, as predicted. $\diamond$

While this simple example is very special in the sense that in practice one would usually apply numerical quadrature rules to integrals whose exact value is unknown, it is, nevertheless, instructive as it provides a recipe for calculating an integral to within a prescribed accuracy, however small it may be, regardless of whether the exact value is known.

3 Global polynomial approximation

In Section 1 we considered the problem of interpolating a function by a polynomial of a certain degree. Here we shall discuss other types of approximation by polynomials, the overall objective being to find the polynomial of degree n which provides the 'best approximation', from $\mathcal{P}_n$, to a given function f in a sense that will be made precise below.

3.1 Normed linear spaces

In order to be able to talk about 'best approximation' in a rigorous manner we shall introduce the concept of *norm*; this will allow us to compare various approximations quantitatively and select the one that has the smallest approximation error. Let $\mathcal{V}$ denote a linear space over $\mathbf{R}$, the field of real numbers. A non-negative function $\|\cdot\|$ defined on $\mathcal{V}$ whose value at $f \in \mathcal{V}$ is denoted $\|f\|$ is called a **norm** on $\mathcal{V}$ if it satisfies the following axioms:

- $\|f\| = 0$ if and only if $f = 0$;
- $\|\lambda f\| = |\lambda|\|f\|$ for all $\lambda \in \mathbf{R}$, and all f in $\mathcal{V}$;
- $\|f + g\| \leq \|f\| + \|g\|$ for all f and g in $\mathcal{V}$; (the triangle inequality).

A linear space $\mathcal{V}$, equipped with a norm, is called a *normed linear space*.

Example 1 *The linear space $\mathbf{R}^n$ is a normed linear space with norm $\|\cdot\|_*$, defined, for $x = (x_1, ..., x_n)^T$, by*

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

Example 2 *The linear space $\mathcal{M}_n$ of all $n \times n$ matrices is a normed linear space with the norm*

$$\|A\|_\star = \max_{x \in \mathbf{R}^n,\, x \neq 0} \frac{\|Ax\|_*}{\|x\|_*},$$

where $\|\cdot\|_*$ denotes the norm defined in Example 1; this matrix norm is said to be induced by the vector norm $\|x\|_*$.

Example 3 *The set $C[a,b]$ of real-valued functions f, defined and continuous on the closed interval $[a,b]$, is a normed linear space with norm*

$$\|f\|_\infty = \max_{x\in[a,b]} |f(x)|.$$

The norm $\|\cdot\|_\infty$ is called the $\mathbf{L^\infty}$ **norm**, *the* **maximum norm** *or, simply, the* ∞**-norm**.

Example 4 *Suppose that w is a real-valued function, defined and continuous on the closed interval (a,b), and that w is positive on (a,b). The set $L^2_w(a,b)$ of real-valued functions f defined on (a,b) and such that $w(x)|f(x)|^2$ is integrable on (a,b) is a normed linear space, equipped with the norm*

$$\|f\|_2 = \left(\int_a^b w(x)|f(x)|^2\,\mathrm{d}x\right)^{1/2},$$

(with the convention that any two functions that are equal on (a,b), except perhaps on a set of measure zero[9] *are identified; $\int_a^b \ldots dx$ in the definition of the norm should be understood as a Lebesgue integral). The norm $\|\cdot\|_2$ is called the* $\mathbf{L^2}$ **norm** *or, simply, the* 2**-norm**. *The function w is called a* **weight function**[10]. *When $w(x)\equiv 1$ on (a,b) we shall simply write $L^2(a,b)$ instead of $L^2_w(a,b)$.*

In Section 3.2 we shall consider the question of constructing the polynomial of best approximation to a function f in the ∞-norm, while Section 3.3 is devoted to best approximation in the 2-norm.

3.2 Approximation in the L^∞ norm

In this section we shall be dealing with the normed linear space $C[a,b]$ of continuous real-valued functions defined on the closed interval $[a,b]$, equipped with the norm $\|\cdot\|_\infty$ (c.f. Example 3). We consider the following approximation problem.

(**A**) Given that $f\in C[a,b]$, find $p_n\in\mathcal{P}_n$ such that

$$\|f-p_n\|_\infty = \inf_{q\in\mathcal{P}_n}\|f-q\|_\infty;$$

such a polynomial p_n is called a **polynomial of best approximation of degree n to the function f in the ∞-norm**.

[9] We say that a set $S\subset[a,b]$ is of *measure zero* if for every $\epsilon>0$ there exists a sequence of non-empty closed intervals $[\alpha_i,\beta_i]\subset[a,b]$, $i=1,2,\ldots$, such that $S\subset\bigcup_{i\geq1}[\alpha_i,\beta_i]$ and $\sum_{i\geq1}(\beta_i-\alpha_i)<\epsilon$.

[10] Our hypotheses on the weight function can be substantially relaxed: it suffices to assume that

a) $w(x)\geq 0$ on the interval $[a,b]$ and $w(x)>0$ for all $x\in[a,b]$, except perhaps for a subset of $[a,b]$ of measure zero (e.g. $w(x)>0$ for all x in (a,b));

b) w is integrable on $[a,b]$; namely, $\int_a^b w(x)\mathrm{d}x<\infty$.

The existence and uniqueness of a polynomial of best approximation for a function $f \in C[a,b]$ in the ∞-norm is ensured by the next theorem.

Theorem 4 *Given that f is a continuous real-valued function defined on the closed interval $[a,b]$, there exists a unique polynomial $p_n \in \mathcal{P}_n$ such that $\|f - p_n\|_\infty = \inf_{q\in\mathcal{P}_n} \|f - q\|_\infty$.*

The uniqueness of the polynomial of best approximation will be proved later on in the section, as a consequence of Theorem 6.

PROOF (OF THEOREM 4): *Existence.* In order to prove the existence of a polynomial of best approximation to a function $f \in C[a,b]$, let us define $d := \inf_{q\in\mathcal{P}_n} \|f - q\|_\infty$. Since $0 \in \mathcal{P}_n$, $d \le \|f - 0\|_\infty = \|f\|_\infty$. According to the definition of d, for every $m \ge 1$ there exists $q_m \in \mathcal{P}_n$ such that

$$\|f - q_m\|_\infty \le d + 1/m; \tag{24}$$

thus, by the triangle inequality, $\|q_m\|_\infty \le 2\|f\|_\infty + 1$, which implies that $|q_m(x)| \le 2\|f\|_\infty + 1$ for each $x \in [a,b]$ and, in particular, choosing $(n+1)$ distinct points x_i, $i = 0, ..., n$, with $a \le x_0 < x_1 < ... < x_n \le b$,

$$\max_{0\le i\le n} |q_m(x_i)| \le 2\|f\|_\infty + 1. \tag{25}$$

Let c_m denote the column-vector of coefficients of the polynomial q_m, and consider the $(n+1)\times(n+1)$ matrix $A = (x_i^j)$, $i, j = 0, ..., n$. Then (25) can be rewritten as

$$\|Ac_m\|_* \le 2\|f\|_\infty + 1,$$

where $\|\cdot\|_*$ is the vector norm defined in Example 1. Since the points x_i are distinct the matrix A is non-singular [11] with inverse matrix A^{-1}; thence, recalling the definition of the matrix norm $\|\cdot\|_\star$ from Example 2,

$$\|c_m\|_* = \|A^{-1}Ac_m\|_* \le \|A^{-1}\|_\star \|Ac_m\|_* \le \|A^{-1}\|_\star (2\|f\|_\infty + 1),$$

which means that $(c_m)_{m\ge1}$ is a bounded sequence in $\mathbf{R}^{n+1}$. Consequently, by the Bolzano–Weierstrass Theorem[12], $(c_m)_{m\ge1}$ has a convergent subsequence $(c_{m_j})_{j\ge1}$. Let us define $\hat{c} = \lim_{j\to\infty} c_{m_j}$ and consider the polynomial $\hat{q}$ in $\mathcal{P}_n$ whose coefficients are the entries of the vector $\hat{c}$; then $\lim_{j\to\infty} \|q_{m_j} - \hat{q}\|_\infty = 0$. Passing to the limit in (24) it follows that $\|f - \hat{q}\|_\infty \le d$, whence $\|f - \hat{q}\|_\infty = d = \inf_{q\in\mathcal{P}_n} \|f - q\|_\infty$ and $\hat{q}$ is a best approximation p_n. □

[11] The matrix A is called the Vandermonde matrix; its determinant, the Vandermonde determinant, $\det(A) = \prod_{0\le i<j\le n}(x_j - x_i)$; since, by assumption, $x_i \ne x_j$ for $i \ne j$, $\det(A) \ne 0$.

[12] **Bolzano – Weierstrass Theorem** (Bernhard Bolzano (1781–1848); Karl Theodor Wilhelm Weierstrass (1815–1897)): Given that $(c_m)_{m\ge1}$ is a bounded sequence in $\mathbf{R}^n$, it has a subsequence $(c_{m_j})_{j\ge1}$ convergent in $\mathbf{R}^n$.

Theorem 4 implies that the *infimum* over $q \in \mathcal{P}_n$ appearing in (**A**) is actually achieved and may be replaced by the *minimum*, $\min_{q \in \mathcal{P}_n}$. Thus, writing the polynomial $q \in \mathcal{P}_n$ in the form

$$q(x) = c_n x^n + ... + c_0,$$

we want to choose the coefficients c_j, $j = 0, ..., n$, to minimise the function

$$\begin{aligned} E(c_0, ..., c_n) &= \|f - q\|_\infty \\ &= \max_{x \in [a,b]} |f(x) - c_n x^n - ... - c_0|. \end{aligned}$$

Since the polynomial of best approximation is to *minimise* (over $q \in \mathcal{P}_n$) the *maximum* absolute value of the error $f(x) - q(x)$ (over $x \in [a, b]$) the polynomial of best approximation in the maximum norm is often referred to as the the **minimax** polynomial. We shall, too, adopt this terminology.

Let us consider a simple example.

Example 5 *Suppose that $f \in C[0, 1]$ and that f is strictly monotonic increasing on $[0, 1]$. We wish to find the minimax polynomial p_0 of degree zero for f on $[0, 1]$. This polynomial will be of the form $p_0(x) = c_0$, and we need to determine c_0 so that*

$$\|f - p_0\|_\infty = \max_{x \in [0,1]} |f(x) - c_0|$$

is minimal. Since f is monotonic increasing $f(x) - c_0$ attains its minimum at $x = 0$ and its maximum at $x = 1$; therefore, $|f(x) - c_0|$ reaches its maximum value at one of the end-points, i.e.

$$E(c_0) = \max_{x \in [0,1]} |f(x) - c_0| = \max\{|f(0) - c_0|, |f(1) - c_0|\}.$$

Clearly,

$$E(c_0) = \begin{cases} f(1) - c_0 & \text{if } c_0 < \frac{1}{2}[f(0) + f(1)] \\ c_0 - f(0) & \text{if } c_0 \geq \frac{1}{2}[f(0) + f(1)]. \end{cases}$$

Drawing the graph of the function $c_0 \mapsto E(c_0)$ shows that the minimum is attained when $c_0 = \frac{1}{2}[f(0) + f(1)]$. Consequently, the minimax polynomial of degree zero to the function f on the interval $[0, 1]$ is

$$p_0(x) \equiv \frac{1}{2}[f(0) + f(1)].$$

This example shows that the minimax approximation of degree zero has the property that it attains the maximum error at two points, with the error being negative at one point and positive at the other. We shall prove that a property of this kind holds in general: the precise formulation of the general result is given in Theorem 6 which is, due to the oscillating nature of the approximation error, usually referred to as the Oscillation Theorem. This theorem gives a complete characterisation of the minimax polynomial and provides a method for its construction. We begin with the following preliminary result.

Theorem 5 (**De la Valée Poussin's Theorem**[13]) *Let $f \in C[a,b]$ and $r \in \mathcal{P}_n$. Given $(n+2)$ points $x_0 < ... < x_{n+1}$ in the interval $[a,b]$, suppose that*

$$\text{sign}\{[f(x_i) - r(x_i)](-1)^i, \quad i = 0, ..., n+1\} = constant,$$

that is, in passing from a point x_i to the next point x_{i+1} the quantity $f(x) - r(x)$ changes sign. Then

$$\min_{q \in \mathcal{P}_n} \|f - q\|_\infty \geq \mu := \min_{i=0,...,n+1} |f(x_i) - r(x_i)|. \tag{26}$$

PROOF: For the case $\mu = 0$ the assertion of the theorem is obvious, so let us assume that $\mu > 0$. Suppose that (26) is not true; then, for the minimax polynomial approximation $p_n \in \mathcal{P}_n$ to the function f, we have that

$$\|f - p_n\|_\infty = \min_{q \in \mathcal{P}_n} \|f - q\|_\infty < \mu.$$

Now

$$\text{sign}[r(x) - p_n(x)] = \text{sign}\{[r(x) - f(x)] - [p_n(x) - f(x)]\}.$$

At the points x_i, the first term exceeds in absolute value the second; therefore $\text{sign}[r(x_i) - p_n(x_i)] = \text{sign}[r(x_i) - f(x_i)]$. Hence the polynomial $r - p_n$, of degree n or less, changes sign $(n+1)$ times. This is a contradiction. □

Now we are ready to state and prove the main result of this section.

Theorem 6 (**The Oscillation Theorem**) *Suppose that $f \in C[a,b]$. For $r \in \mathcal{P}_n$ to be a minimax polynomial approximation to f over $[a,b]$ it is necessary and sufficient that there exists a sequence of $(n+2)$ points x_j, where $a \leq x_0 \leq ... \leq x_{n+1} \leq b$, such that*

$$f(x_i) - r(x_i) = \alpha(-1)^i \|f - r\|_\infty, \qquad i = 0, ..., n+1,$$

where $\alpha = 1$ (or $\alpha = -1$) simultaneously for all i.

The points $x_0, ..., x_{n+1}$ which satisfy the conditions of the theorem are called **critical points**.

PROOF: *Sufficiency.* Let L denote the quantity $\|f - r\|_\infty$, and define $E_n(f) = \min_{q \in \mathcal{P}_n} \|f - q\|_\infty$. Applying (26) it follows that $L = \mu \leq E_n(f)$. However, by the definition of $E_n(f)$ we also have that $E_n(f) \leq \|f - r\|_\infty = L$. Hence $E_n(f) = L$, and the given polynomial r is a minimax polynomial.

Necessity. Suppose that the given polynomial r is a minimax polynomial. Let us denote by y_1 the lower bound of points $x \in [a,b]$ at which $|f(x) - r(x)| = L$; the existence of such a point follows from the definition of L. Because $f - r$ is a continuous function on $[a,b]$, we have $|f(y_1) - r(y_1)| = L$. We can assume, without restricting generality, that $f(y_1) - r(y_1) = L$. Let us denote by y_2 the lower bound of all points $x \in (y_1, b]$ at which $f(x) - r(x) = -L$, and denote, in succession, by

[13] Jean Charles de la Valée Poussin (1866–1962)

y_{k+1} the lower bound of points $x \in (y_k, b]$ at which $f(x) - r(x) = (-1)^k L$, etc. Due to the continuity of the function $f - r$, we have, for each k, that

$$f(y_{k+1}) - r(y_{k+1}) = (-1)^k L.$$

Let us continue this process up to $y_m = b$, or y_m such that $|f(x) - r(x)| < L$ for $y_m < x \leq b$. If $m \geq n + 2$, the proof is complete. Let us therefore assume that $m < n + 2$ and show that this leads to contradiction.

Since $f - r$ is continuous, for each k, $1 < k \leq m$, there exists a point z_{k-1} such that $|f(x) - r(x)| < L$ for $z_{k-1} \leq x < y_k$. We define $z_0 = a$ and $z_m = b$. According to our construction, there exist points in the intervals $[z_{i-1}, z_i]$, $i = 1, ..., m$, at which $f(x) - r(x) = (-1)^{i-1} L$ (such are the points y_i, for example), and there is no point x in $[z_{i-1}, z_i]$ where $f(x) - r(x) = (-1)^i L$. We define

$$v(x) = \prod_{j=1}^{m-1} (z_j - x), \quad \text{and} \quad r(x; \epsilon) = r(x) + \epsilon\, v(x), \;\; \epsilon > 0,$$

and consider the behaviour of the difference

$$f(x) - r(x; \epsilon) = f(x) - r(x) - \epsilon\, v(x)$$

on the intervals $[z_{j-1}, z_j]$. Take, for example, the interval $[z_0, z_1]$. On $[z_0, z_1)$ we have $v(x) > 0$ and therefore

$$f(x) - r(x; \epsilon) \leq L - \epsilon\, v(x) < L.$$

At the same time, $f(x) - r(x) > -L$ on $[z_0, z_1]$, and so for ϵ sufficiently small, say, for

$$\epsilon < \epsilon_1 = \frac{\min_{x \in [z_0, z_1]} |f(x) - r(x) + L|}{\max_{x \in [z_0, z_1]} |v(x)|},$$

we have that $f(x) - r(x; \epsilon) > -L$ for all x in $[z_0, z_1)$. Furthermore,

$$|f(z_1) - r(z_1; \epsilon)| = |f(z_1) - r(z_1)| < L.$$

Therefore $|f(x) - r(x; \epsilon)| < L$ for all $x \in [z_0, z_1]$, for ϵ sufficiently small. Arguing in the same manner on the other intervals $[z_{j-1}, z_j]$, $j = 2, ..., m$, we can choose ϵ_0 such that

$$|f(x) - r(x; \epsilon_0)| < L, \qquad \text{for } x \in \textstyle\bigcup_{j=1}^{m} [z_{j-1}, z_j] = [a, b].$$

Since, by hypothesis, $m < n + 2$, it follows that $m - 1 < n + 1$, and therefore $v \in \mathcal{P}_n$; consequently, $r(x; \epsilon_0) \in \mathcal{P}_n$. Thus we have constructed a polynomial $r(x; \epsilon_0) \in \mathcal{P}_n$ such that

$$\|f - r(\cdot; \epsilon_0)\|_\infty < L = \|f - r\|_\infty,$$

which contradicts the assumption that r is a polynomial of best approximation to the function f on the interval $[a, b]$ and, simultaneously, $m < n + 2$. Thus, to summarise, assuming that $m < n + 2$ we arrived at a contradiction, which implies

that $m \geq n+2$. Hence we have proved the existence of m critical points $y_1,...,y_m$, $m \geq n+2$; choosing $n+2$ consecutive points from this set and naming them $x_0,...,x_{n+1}$ we obtain the desired result, and the proof is complete. □

Equipped with the Oscillation Theorem, we are now ready to complete the proof of Theorem 4 and to show the uniqueness of the minimax polynomial.

PROOF OF THEOREM 4: *Uniqueness.* Suppose that there are two minimax polynomials, p_n and $\hat{p}_n$, to the function $f \in C[0,1]$ on the interval $[a,b]$:

$$p_n \neq \hat{p}_n, \qquad \|f - p_n\|_\infty = \|f - \hat{p}_n\|_\infty = E_n(f),$$

where, as in the proof of the Oscillation Theorem, we have used the notation $E_n(f) = \min_{q \in \mathcal{P}_n} \|f - q\|_\infty$. This implies, by the triangle inequality, that

$$\begin{aligned} \|f - \frac{1}{2}(p_n + \hat{p}_n)\|_\infty &= \|\frac{1}{2}(f - p_n) + \frac{1}{2}(f - \hat{p}_n)\|_\infty \\ &\leq \frac{1}{2}\|f - p_n\|_\infty + \frac{1}{2}\|f - \hat{p}_n\|_\infty = E_n(f). \end{aligned}$$

Therefore, $\frac{1}{2}(p_n + \hat{p}_n)$ is also a minimax polynomial approximation to f on $[a,b]$. By the Oscillation Theorem, there exist $(n+2)$ critical points x_i, $i = 0, ..., n+1$, corresponding to this polynomial at which

$$\left| f(x_i) - \frac{1}{2}(p_n(x_i) + \hat{p}_n(x_i)) \right| = E_n(f), \qquad i = 0, ..., n+1.$$

Equivalently,

$$|[f(x_i) - p_n(x_i)] + [f(x_i) - \hat{p}_n(x_i)]| = 2E_n(f). \tag{27}$$

Since

$$|f(x_i) - p_n(x_i)| \leq \max_{x \in [a,b]} |f(x) - p_n(x)| = \|f - p_n\|_\infty = E_n(f),$$

and, analogously,

$$|f(x_i) - \hat{p}_n(x_i)| \leq E_n(f),$$

it follows from (27) that we must have

$$f(x_i) - p_n(x_i) = f(x_i) - \hat{p}_n(x_i), \qquad i = 0, ..., n+1.$$

Thus $(p_n - \hat{p}_n)(x_i) = 0$, $i = 0, ..., n+1$. Given that $p_n - \hat{p}_n$ is a polynomial of degree n or less, it can have more than n zeros only if it is identically zero, i.e. $p_n(x) \equiv \hat{p}_n(x)$. However this contradicts our initial hypothesis that p_n and $\hat{p}_n$ are distinct. Hence there is a unique minimax polynomial approximation in $\mathcal{P}_n$ to the function $f \in C[a,b]$ on the interval $[a,b]$. □

In order to demonstrate the application of the Oscillation Theorem to a specific problem we consider the following example.

Exercise 5 *Suppose that $f \in C[a,b]$ and that f is a convex function on $[a,b]$ such that $f'(x)$ exists at all x in (a,b). Describe a method for constructing the minimax polynomial approximation $p_1 \in \mathcal{P}_1$ of degree 1 to f on the interval $[a,b]$.*

SOLUTION: We seek p_1 in the form $p_1(x) = c_1 x + c_0$. Due to the convexity of f the difference $f(x) - (c_1 x + c_0)$ can only have one interior extremum point. Therefore the end-points of the interval, a and b, are critical points. Let us denote by d the third critical point whose location inside (a,b) remains to be determined. By the Oscillation Theorem we have the equations:

$$\begin{aligned} f(a) - (c_1 a + c_0) &= \alpha L, \\ f(d) - (c_1 d + c_0) &= -\alpha L, \\ f(b) - (c_1 b + c_0) &= \alpha L, \end{aligned}$$

where $L = \max_{x\in[a,b]} |f(x) - p_1(x)|$, and $\alpha = 1$ or $\alpha = -1$. We have a total of only three equations to determine the unknowns d, c_1, c_0, L and α.

Subtracting the first equation from the third, we get $f(b) - f(a) = c_1(b-a)$, whereby $c_1 = [f(b) - f(a)]/(b-a)$. In order to find the remaining unknowns it should be borne in mind that the point d is an extremum point of the difference $f(x) - (c_1 x + c_0)$. This additional condition supplies sufficient information to enable us to determine d; in particular as f is differentiable on (a,b), we deduce that d can be found from the equation $f'(d) - c_1 = 0$. Now c_0 can be determined by adding the second equation to the first. Having calculated both c_1 and c_0 we insert them into the first equation to obtain αL; finally, $L = |\alpha L|$, while $\alpha = \text{sign}(\alpha L)$. $\diamond$

Exercise 6 *Construct the minimax polynomial approximation $p_1(x)$ of degree 1 for $f(x) = \tan^{-1} x$ on the interval $[0,1]$.*

SOLUTION: Noting that $f : x \mapsto \tan^{-1} x$ is a continuous, convex function on the interval $[0,1]$, and $f'(x)$ exists at all $x \in (0,1)$, we seek $p_1(x) = c_1 x + c_0$ following the guidelines of the previous exercise. The set of critical points is $\{0, d, 1\}$ and, by the Oscillation Theorem,

$$\begin{aligned} f(0) - p_1(0) &= \alpha L, \\ f(d) - p_1(d) &= -\alpha L, \\ f(1) - p_1(1) &= \alpha L, \end{aligned}$$

where $L = \max_{x\in[0,1]} |f(x) - p_1(x)|$ and $\alpha = \pm 1$. Furthermore, since d is a point of internal extremum of $f(x) - (c_1 x + c_0)$, we also have that

$$f'(d) - c_1 = 0.$$

We solve this set of four equations for c_0, c_1, d and αL, and then take $\alpha = \text{sign}(\alpha L)$. Thus,

$$\begin{aligned} c_0 &= -\alpha L, \\ c_0 &= \frac{1}{2}(\tan^{-1} d - c_1 d), \\ c_1 &= \tan^{-1} 1, \\ c_1 &= \frac{1}{1+d^2}. \end{aligned}$$

From the last two equations we have that

$$c_1 = 0.78540 \qquad \text{and} \qquad d = 0.52272.$$

Hence

$$c_0 = 0.03556,$$

and therefore

$$\alpha L = -0.03556,$$

so that $\alpha = -1$ and $L = 0.03556$. Consequently,

$$p_1(x) = 0.78540x + 0.03556$$

is the minimax polynomial of degree 1 for the function $f : x \mapsto \tan^{-1} x$ on the interval $[0, 1]$, and the error of the approximation is $L = 0.03556$. $\diamond$

3.2.1 Minimax approximation to x^{n+1}

There are very few functions for which it is possible to write down in simple closed form the polynomial of best approximation in the maximum norm. One such problem of practical importance concerns the approximation of a power of x by a polynomial of lower degree. The minimax approximation in this case is given in terms of Chebyshev polynomials. The Chebyshev polynomial of degree n is defined by

$$T_n(x) = \cos(n \cos^{-1} x), \qquad n = 0, 1, ...$$

Despite its unusual form, T_n is a polynomial in disguise: for example, $T_0(x) \equiv 1$, $T_1(x) = x$, and so on. In order to show that T_n is a polynomial for all $n \geq 0$, let us recall the standard trigonometric identity

$$\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta,$$

and set $\theta = \cos^{-1} x$ to obtain the recurrence relation

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \qquad n = 1, 2, ...$$

Since $T_0(x)$ and $T_1(x)$ have already been shown to be polynomials, we deduce from this recurrence relation, by induction, that $T_n(x)$ is a polynomial of degree n for each $n \geq 0$. A list of the first seven Chebyshev polynomials is given in Example 9 in Section 3.3.3.

The proof of the next lemma is straightforward and is left as an exercise (see also Exercise 2).

Lemma 3 *The Chebyshev polynomials satisfy the following properties:*

a) $T_n(x)$ *is a polynomial of degree* n*, with leading coefficient* $2^{n-1}x^n$*.*

b) $T_n(x)$ *is an even function if* n *is even, and an odd function if* n *is odd.*

c) The zeros of $T_n(x)$ *are at*

$$x_j = \cos\frac{(j-\frac{1}{2})\pi}{n}, \qquad j = 1, ..., n.$$

They are all real and distinct, and lie in $(-1, 1)$.

d) $|T_n(x)| \leq 1$.

e) $T_n(x) = \pm 1$, *alternately at the* $(n+1)$ *points* $x_k = \cos k\pi/n$, $k = 0, 1, ..., n$.

According to the Oscillation Theorem, the minimax polynomial approximation to $f(x) = x^{n+1}$ over the interval $[-1, 1]$ by a polynomial of degree n is the polynomial $p_n(x)$ such that the difference $f(x) - p_n(x)$ attains its greatest magnitude with alternating signs at a sequence of $(n+2)$ points in $[-1, 1]$. Let us define

$$p_n(x) = x^{n+1} - 2^{-n}T_{n+1}(x);$$

we shall prove that $p_n(x)$ is the minimax approximation of degree n to x^{n+1} on the interval $[-1, 1]$. Clearly $p_n \in \mathcal{P}_n$. Since

$$x^{n+1} - p_n(x) = 2^{-n}T_{n+1}(x),$$

the difference $x^{n+1} - p_n(x)$ does not exceed 2^{-n} in the interval $[-1, 1]$, and attains this value with alternating signs at the $(n+2)$ points $x_j = \cos j\pi/(n+1)$, $j = 0, ..., n+1$. Therefore, by the Oscillation Theorem, $p_n(x)$ is the (unique) minimax polynomial approximation from $\mathcal{P}_n$ to the function x^{n+1} over $[-1, 1]$.

3.3 Approximation in the L^2 norm

Best approximation in the 2-norm is closely related to the notion of orthogonality and this in turn relies on the concept of inner product.

3.3.1 Inner product spaces

Let $\mathcal{V}$ be a real linear space. A real-valued function $(\cdot, \cdot)$ defined on the cartesian product $\mathcal{V} \times \mathcal{V}$ is called an **inner product** on $\mathcal{V}$ if it satisfies the following axioms:

- $(f + g, h) = (f, h) + (g, h)$ for all f, g and h in $\mathcal{V}$;
- $(\lambda f, g) = \lambda(f, g)$ for all λ in $\mathbf{R}$, and f, g in $\mathcal{V}$;
- $(f, g) = (g, f)$ for all f and g in $\mathcal{V}$;
- $(f, f) > 0$ if $f \neq 0$, $f \in \mathcal{V}$.

A linear space with an inner product is called an **inner product space**. If $(f, g) = 0$ for f and g in $\mathcal{V}$, we say that f is **orthogonal** to g. For f in $\mathcal{V}$, we define

$$\|f\| := (f, f)^{1/2}.$$

It is left to the reader to show that, with such a definition of $\|\cdot\|$,

$$|(f, g)| \leq \|f\| \, \|g\|, \qquad f,\, g \in \mathcal{V} \text{ (the Cauchy-Schwarz inequality)}^{14},$$

and, therefore,

$$\|f + g\| \leq \|f\| + \|g\|, \;\; f,\, g \in \mathcal{V}.$$

Consequently $\|\cdot\|$ is a norm on $\mathcal{V}$ (induced by the inner product $(\cdot, \cdot)$, and $\mathcal{V}$ is a normed linear space.

Example 6 *The n-dimensional euclidean space* $\mathbf{R}^n$ *is an inner product space with*

$$(x, y) = \sum_{i=1}^{n} x_i y_i,$$

where $x = (x_1, ..., x_n)$ *and* $y = (y_1, ..., y_n)$.

Example 7 *The set* $C[a, b]$ *of continuous real-valued functions defined on the closed interval* $[a, b]$ *is an inner product space with*

$$(f, g) = \int_0^1 w(x) f(x)\, g(x)\, \mathrm{d}x,$$

where w *is a weight function, defined and continuous on* $[0, 1]$ *and positive on* $(0, 1)$. *Clearly, this inner product induces the 2-norm (see Example 4) in the sense that* $\|f\|_2 = (f, f)^{1/2}$.

Having introduced the concept of inner product, we are now ready to consider best approximation in the 2-norm.

3.3.2 Best approximation in the L^2 norm

Let w be a real-valued function, defined and continuous on a closed interval $[a, b]$, and suppose that w is positive on (a, b); w will be referred to as a **weight function**. Let $L^2_w(a, b)$ denote the set of all real-valued functions f defined on (a, b) such that $w(x)|f(x)|^2$ is integrable on (a, b); the set $L^2_w(a, b)$ is equipped with the 2-norm defined, as in Example 4, by

$$\|f\|_2 = \left(\int_a^b w(x) |f(x)|^2 \, \mathrm{d}x \right)^{1/2}.$$

The problem of best approximation in the 2-norm can be formulated as follows:

[14] Hint: $0 \leq \|\lambda f + g\|_2^2 = \lambda^2 \|f\|_2^2 + 2\lambda(f, g) + \|g\|_2^2$, for all $\lambda \in \mathbf{R}$; the quadratic polynomial on the right is non-negative for all real λ if and only if $(2(f, g))^2 - 4\|f\|_2 \|g\|_2 \leq 0$.

(**B**) Given that $f \in L^2_w(a,b)$, find $p_n \in \mathcal{P}_n$ such that

$$\|f - p_n\|_2 = \inf_{q \in \mathcal{P}_n} \|f - q\|_2;$$

such p_n is called a **polynomial of best approximation of degree n to the function f in the 2-norm**.

The next theorem ensures the existence and uniqueness of the polynomial of best approximation to a function f of a given degree in the 2-norm.

Theorem 7 *Given that $f \in L^2_w(a,b)$, there exists a unique polynomial $p_n \in \mathcal{P}_n$ such that $\|f - p_n\|_2 = \inf_{q \in \mathcal{P}_n} \|f - q\|_2$.*

The proof of this result will be given in Section 3.3.3. Let us illustrate the general approach to problem (**B**) by a simple example. Suppose that we wish to construct the polynomial approximation $p_n \in \mathcal{P}_n$ to a function f on the interval $[0,1]$; for simplicity, we shall assume that the weight function $w(x) \equiv 1$. Writing the polynomial

$$p_n(x) = c_n x^n + ... + c_1 x + c_0,$$

we want to choose the coefficients c_j so as to minimise the 2-norm of the error, $e_n = f - p_n$,

$$\|e_n\|_2 = \|f - p_n\|_2 = \left(\int_0^1 |f(x) - p_n(x)|^2 \, \mathrm{d}x \right)^{1/2}.$$

Since the 2-norm is positive, this problem is equivalent to the minimisation of the square of the norm; thus we shall minimise the expression

$$\begin{aligned} E(c_0, c_1, ..., c_n) &= \int_0^1 [f(x) - p_n(x)]^2 \, \mathrm{d}x \\ &= \int_0^1 [f(x)]^2 \, \mathrm{d}x - 2 \sum_{j=0}^{n} c_j \int_0^1 f(x) x^j \, \mathrm{d}x \\ &\qquad + \sum_{j=0}^{n} \sum_{k=0}^{n} c_j c_k \int_0^1 x^{j+k} \, \mathrm{d}x. \end{aligned}$$

At the unique minimum, the partial derivatives of E with respect to the c_j, $j = 0, ..., n$, are equal to zero. This leads to a system of $(n+1)$ linear equations for the coefficients c_0,...,c_n:

$$\sum_{k=0}^{n} M_{jk} c_k = b_j, \qquad j = 0, ..., n, \tag{28}$$

where

$$\begin{aligned} M_{jk} &= \int_0^1 x^{j+k} \, \mathrm{d}x = \frac{1}{j+k+1}, \\ b_j &= \int_0^1 f(x) x^j \, \mathrm{d}x. \end{aligned}$$

Equivalently, recalling that the inner product associated with the 2-norm (in the case of $w(x) \equiv 1$) is defined by

$$(g, h) = \int_0^1 g(x)h(x)\,\mathrm{d}x,$$

M_{jk} and b_j can be written as

$$M_{jk} = (x^k, x^j), \qquad b_j = (f, x^j). \tag{29}$$

By solving the system of linear equations (28) we obtain the coefficients of the polynomial of best approximation of degree n to the function f in the 2-norm on the interval $[0, 1]$. We can proceed in the same manner on any interval $[a, b]$ and any positive weight-function w.

Exercise 7 *Given that $f(x) = x^2$ for $x \in [0, 1]$, find the polynomial p_1 of degree 1 of best approximation to f in the 2-norm on the interval $[0, 1]$ assuming that the weight function is $w(x) \equiv 1$.*

SOLUTION: We seek $p_1(x) = c_1 x + c_0$ such that

$$E(c_0, c_1) = \int_0^1 [x^2 - (c_1 x + c_0)]^2\,\mathrm{d}x \to \text{minimum}.$$

At the minimum, we have that $\frac{\partial E}{\partial c_0} = 0$ and $\frac{\partial E}{\partial c_1} = 0$; therefore,

$$\begin{aligned} \int_0^1 2(x^2 - (c_1 x + c_0)) \cdot (-1)\,\mathrm{d}x &= 0, \\ \int_0^1 2(x^2 - (c_1 x + c_0)) \cdot (-x)\,\mathrm{d}x &= 0. \end{aligned}$$

Upon evaluating the integrals we arrive at the following system of linear equations:

$$\begin{aligned} c_0 + \frac{1}{2}c_1 &= \frac{1}{3} \\ \frac{1}{2}c_0 + \frac{1}{3}c_1 &= \frac{1}{4}. \end{aligned}$$

Solving this gives $c_0 = -\frac{1}{6}$ and $c_1 = 1$ and therefore,

$$p_1(x) = x - \frac{1}{6}$$

is the required polynomial of best approximation. $\diamond$

Returning to the general discussion concerning the solution of the linear system (28), we see that we have to invert the matrix $M = (M_{jk})$ with $(n + 1)$ rows and columns[15]; this is quite a simple task for small values of n (such as $n = 1, 2, 3$; indeed, we encountered the 2×2 Hilbert matrix, corresponding to $n = 1$, in the previous exercise), but for larger n we need a more effective approach: in the next section we shall discuss an alternative technique, based on the use of orthogonal polynomials.

[15]The matrix $M = (M_{jk})$ with $M_{jk} = 1/(j+k+1)$ is called the Hilbert matrix. It is notoriously difficult to invert because it is close to being singular. For example, for $n = 10$ when the matrix is of size 11×11, the smallest eigenvalue is approximately 1.9×10^{-13}.

3.3.3 Orthogonal polynomials

In the previous section we described a method for constructing the polynomial of best approximation $p_n \in \mathcal{P}_n$ to a function f in the 2-norm; it was based on seeking p_n as a linear combination of the polynomials x^j, $j = 0, ..., n$, which form a basis for the linear space $\mathcal{P}_n$. The approach was not entirely satisfactory because it gave rise to a system of linear equations with a full matrix that was difficult to invert. The central idea of the alternative approach that will be described in this section is to expand p_n in terms of a different basis, chosen so that the resulting system of linear equations has a diagonal matrix and solving this linear system is therefore a trivial exercise. Of course, the non-trivial part of the problem is then to find a suitable basis for $\mathcal{P}_n$ that achieves this goal. The expression for M_{jk} in (29) gives us a clue how to proceed.

Suppose that $\phi_j(x)$, $j = 0, ..., n$, form a basis for $\mathcal{P}_n$; let us seek the polynomial of best approximation as

$$p_n(x) = \gamma_0\phi_0(x) + ... + \gamma_n\phi_n(x).$$

Repeating the same process as in the previous section, we arrive at a system of linear equations of the form (22):

$$\sum_{k=0}^{n} M_{jk}\gamma_k = \beta_j, \qquad j = 0, ..., n,$$

where now

$$M_{jk} = (\phi_k, \phi_j), \quad \text{and} \quad \beta_j = (f, \phi_j).$$

Thus, $M = (M_{jk})$ will be a diagonal matrix provided the basis functions $\phi_j(x)$, $j = 0, ..., n$, for the space $\mathcal{P}_n$ are chosen so that $(\phi_k, \phi_j) = 0$, for $j \neq k$; in other words, using the terminology introduced in Section 3.3.1, ϕ_k is required to be orthogonal to ϕ_j for $j \neq k$. This motivates the following definition.

Definition 3 *Given a weight function w, defined and continuous on the interval $[a, b]$ and positive on (a, b), we say that the sequence of polynomials $\phi_j(x)$, $j = 0, 1, ...,$ forms a* **system of orthogonal polynomials** *on the interval (a, b) with respect to w, if each $\phi_j(x)$ is of exact degree j, and if*

$$\begin{aligned} \int_a^b w(x)\phi_j(x)\phi_k(x)\,dx &= 0 \qquad \textit{for all } j \neq k, \\ &\neq 0 \qquad \textit{when } j = k. \end{aligned}$$

We show that a sequence of orthogonal polynomials exists on any interval (a, b) and for any weight function w that is continuous on $[a, b]$ and positive on (a, b), by providing a method of construction.

Let $\phi_0(x) \equiv 1$, and suppose that $\phi_j(x)$ has already been constructed for $j = 0, ..., n$. Then

$$\int_a^b w(x)\phi_j(x)\phi_k(x)\,\mathrm{d}x = 0, \qquad 0 \leq j < k \leq n.$$

Now let us define the polynomial

$$q(x) = x^{n+1} - a_0\phi_0(x) - ... - a_n\phi_n(x),$$

where

$$a_j = \frac{\int_a^b w(x)x^{n+1}\phi_j(x)\,\mathrm{d}x}{\int_a^b w(x)[\phi_j(x)]^2\,\mathrm{d}x}.$$

Then it follows that

$$\begin{aligned}\int_a^b w(x)q(x)\phi_j(x)\,\mathrm{d}x &= \int_a^b w(x)x^{n+1}\phi_j(x)\,\mathrm{d}x \\ &\qquad -a_j\int_a^b w(x)[\phi_j(x)]^2\,\mathrm{d}x, \\ &= 0 \qquad \text{for } 0 \le j \le n,\end{aligned}$$

where we have used the orthogonality of the sequence for $j = 0, 1, ..., n$. Thus, with this choice of the numbers a_j we have ensured that $q(x)$ is orthogonal to all the previous members of the sequence, and $\phi_{n+1}(x)$ can now be defined as any non-zero-constant multiple of $q(x)$. This procedure for constructing a system of orthogonal polynomials is usually referrred to as **Gram-Schmidt orthogonalistion**.

Exercise 8 *Construct a system of orthogonal polynomials* $\{\psi_0, \psi_1, \psi_2\}$ *on the interval* $(0, 1)$ *with respect to the weight function* $w(x) \equiv 1$.

SOLUTION: Given that $w(x) \equiv 1$, the inner product of $L^2_w(0, 1)$ is defined by

$$(u, v) = \int_0^1 u(x)v(x)\,\mathrm{d}x.$$

We put $\psi_0(x) \equiv 1$, and we seek ψ_1 in the form

$$\psi_1(x) = x - c_0\psi_0(x)$$

such that $(\psi_1, \psi_0) = 0$; namely,

$$(x, \psi_0) - c_0(\psi_0, \psi_0) = 0.$$

Hence,

$$c_0 = \frac{(x, \psi_0)}{(\psi_0, \psi_0)} = \frac{1/2}{1} = \frac{1}{2}$$

and therefore,

$$\psi_1(x) = x - \frac{1}{2}\psi_0(x) = x - \frac{1}{2}.$$

By construction, $(\psi_1, \psi_0) = (\psi_0, \psi_1) = 0$.

Now we seek ψ_2 in the form

$$\psi_2(x) = x^2 - (d_1\psi_1(x) + d_0\psi_0(x))$$

such that $(\psi_2, \psi_1) = 0$ and $(\psi_2, \psi_0) = 0$. Thus,

$$\begin{aligned}(x^2, \psi_1) - d_1(\psi_1, \psi_1) - d_0(\psi_0, \psi_1) &= 0,\\ (x^2, \psi_0) - d_1(\psi_1, \psi_0) - d_0(\psi_0, \psi_0) &= 0.\end{aligned}$$

As $(\psi_0, \psi_1) = 0$ and $(\psi_1, \psi_0) = 0$, we have that

$$\begin{aligned}d_1 &= \frac{(x^2, \psi_1)}{(\psi_1, \psi_1)} = 1,\\ d_0 &= \frac{(x^2, \psi_0)}{(\psi_0, \psi_0)} = \frac{1}{3},\end{aligned}$$

and therefore

$$\psi_2(x) = x^2 - x + \frac{1}{6}.$$

Clearly, $(\psi_j, \psi_k) = 0$ for $j \neq k$, $j, k \in \{0, 1, 2\}$, and ψ_k is of exact degree k, $k = 0, 1, 2$, so we have found the required system of orthogonal polynomials on the interval $(0, 1)$. $\diamond$

Example 8 (**Legendre polynomials**[16]) *The polynomials*

$$\begin{aligned}\phi_0(x) &= 1,\\ \phi_1(x) &= x,\\ \phi_2(x) &= x^2 - \frac{1}{3},\\ \phi_3(x) &= x^3 - \frac{3}{5}x\end{aligned}$$

are the first four elements of an orthogonal system on the interval $(-1, 1)$ *with respect to the weight function* $w(x) \equiv 1$.

Example 9 (**Chebyshev polynomials**[17]) *The polynomials*

$$\begin{aligned}T_0(x) &= 1,\\ T_1(x) &= x,\\ T_2(x) &= 2x^2 - 1,\\ T_3(x) &= 4x^3 - 3x,\\ T_4(x) &= 8x^4 - 8x^2 + 1,\\ T_5(x) &= 16x^5 - 20x^3 + 5x,\\ T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1\end{aligned}$$

are the first seven elements of an orthogonal system on the interval $(-1, 1)$ *with respect to the positive weight function* $w(x) = (1 - x^2)^{-1/2}$, $x \in (-1, 1)$. *(This weight function is continuous on* $(-1, 1)$*, but not on* $[-1, 1]$*; see, however, Footnote 10 on page 18.)*

[16] Adrien-Marie Legendre (1752–1833)

[17] P.L. Chebyshev (1821–1894)

We are now ready to prove Theorem 7.

PROOF (OF THEOREM 7): In order to simplify the notation, we recall the definition of the inner product $(\cdot,\cdot)$:

$$(g,h) = \int_a^b w(x)g(x)h(x)\,\mathrm{d}x,$$

and note that the associated 2-norm can be expressed as $\|g\|_2 = (g,g)^{1/2}$. Let $\phi_i(x)$, $i = 0,1,...,n$, be a system of orthogonal polynomials with respect to the weight function w on (a,b). Let us normalise the polynomials ϕ_i by defining a new set of orthogonal polynomials,

$$\psi_i(x) = \frac{\phi_i(x)}{\|\phi_i\|_2}, \qquad i = 1,...,n.$$

Then

$$(\psi_j,\psi_k) = \begin{cases} 1, & j = k \\ 0, & j \neq k. \end{cases}$$

The polynomials $\psi_i(x)$, $i = 1,...,n$, are linearly independent and form a basis for the linear space $\mathcal{P}_n$; therefore, each element $q \in \mathcal{P}_n$ can be expressed as their linear combination,

$$q(x) = \beta_0\psi_0(x) + ... + \beta_n\psi_n(x).$$

Consider the function

$$\begin{aligned} E(\beta_0,...,\beta_n) &= \|f-q\|_2^2 = (f-q,f-q) \\ &= (f,f) - 2(f,q) + (q,q) \\ &= \|f\|_2^2 - 2\sum_{j=0}^{n}\beta_j(f,\psi_j) + \sum_{j=0}^{n}\sum_{k=0}^{n}\beta_j\beta_k(\psi_j,\psi_k) \\ &= \|f\|_2^2 - 2\sum_{j=0}^{n}\beta_j(f,\psi_j) + \sum_{j=0}^{n}\beta_j^2. \end{aligned}$$

Clearly E is a strictly concave quadratic function of its arguments β_0,..., β_n, and therefore it has a unique minimum $(\beta_0^*,...,\beta_n^*) \in \mathbf{R}^{n+1}$. Hence $p_n \in \mathcal{P}_n$, defined by

$$p_n(x) = \beta_0^*\psi_0(x) + ... + \beta_n^*\psi_n(x)$$

is the unique polynomial of best approximation of degree n to the function $f \in L_w^2(a,b)$ in the 2 norm on the interval (a,b). □

The next theorem, in conjunction with the use of orthogonal polynomials, is the key tool for constructing the polynomial of best approximation in the 2-norm.

Theorem 8 *The polynomial $p_n \in \mathcal{P}_n$ is the polynomial of best approximation of degree n to a function $f \in L_w^2(a,b)$ in the 2-norm if and only if the difference $f - p_n$ is orthogonal to every element of $\mathcal{P}_n$, i.e.*

$$\int_a^b w(x)(f(x) - p_n(x))q(x)\,dx = 0 \qquad \textit{for all } q \in \mathcal{P}_n.$$

PROOF: Using the same notation as in the proof of Theorem 7, we note that any polynomial $p_n \in \mathcal{P}_n$, can be written as

$$p_n(x) = \beta_0\psi_0(x) + ... + \beta_n\psi_n(x). \tag{30}$$

Now p_n is a polynomial of best approximation of degree n in the 2-norm to the function $f \in L^2_w(a,b)$ if and only if $\beta_i = \beta_i^*$, $i = 0, ..., n$, where $(\beta_0^*, ..., \beta_n^*)$ is the unique minimum of the function

$$E(\beta_0, ..., \beta_n) = \|f\|_2^2 - 2\sum_{j=0}^{n} \beta_j(f, \psi_j) + \sum_{j=0}^{n} \beta_j^2.$$

But, at the minimum, the partial derivatives of E with respect to the β_i, $i = 0, ..., n$, are equal to 0; therefore,

$$\beta_i^* = (f, \psi_i).$$

Inserting these values into (30) we deduce that p_n is the (unique) polynomial of best approximation to f in the 2-norm if and only if it can be expressed as

$$p_n(x) = \sum_{i=0}^{n} (f, \psi_i)\psi_i(x).$$

Thus if, p_n is a polynomial of best approximation in the 2-norm then

$$(f - p_n, \psi_j) = (f - \sum_{i=0}^{n} (f, \psi_i)\psi_i, \psi_j) = 0, \qquad j = 0, ..., n,$$

and therefore also

$$(f - p_n, q) = 0 \qquad \text{for all } q \in \mathcal{P}_n.$$

Conversely, if $(f - p_n, q) = 0$ for all q in $\mathcal{P}_n$ then also $(f - p_n, \psi_j) = 0$ for all $j = 0, ..., n$. Writing p_n as

$$p_n(x) = \beta_0\psi_0(x) + ... + \beta_n\psi_n(x).$$

it follows that $\beta_i = (f, \psi_i) = \beta_i^*$, and therefore p_n is the polynomial of best approximation in the 2-norm. □

Theorem 8 provides a simple method of determining the polynomial of best approximation p_n to a function $f \in L^2_w(a,b)$ in the 2-norm. First, proceeding as described in the discussion following Definition 3, we construct the system of orthogonal polynomials $\phi_j(x)$, $j = 0, ..., n$, on the interval (a,b) with respect to the weight function w, if this system is not already known. Then we seek p_n as the linear combination

$$p_n(x) = \gamma_0\phi_0(x) + ... + \gamma_n\phi_n(x).$$

By virtue of Theorem 8, the difference $f - p_n$ must be orthogonal to every polynomial of degree n or less, and in particular to each polynomial ϕ_j, $j = 0, ..., n$. Thus

$$\int_a^b w(x) \left[f(x) - \sum_{k=0}^{n} \gamma_k\phi_k(x) \right] \phi_j(x)\,\mathrm{d}x = 0, \qquad j = 0, 1, ..., n.$$

Exploiting the orthogonality of the polynomials ϕ_j, this gives

$$\gamma_j = \frac{\int_a^b w(x) f(x) \phi_j(x)\,\mathrm{d}x}{\int_a^b w(x)[\phi_j(x)]^2\,\mathrm{d}x} \left(= \frac{(f,\phi_j)}{\|\phi_j\|_2^2} \right).$$

Thus, as indicated at the beginning of the section, with this approach to the construction of the polynomial of best approximation in the 2-norm, we obtain the coefficients γ_j explicitly and there is no need to solve a system of linear equations with a full matrix.

Exercise 9 *Construct the polynomial of best approximation of degree* 2 *in the* 2*-norm to the function* $f : x \mapsto \mathrm{e}^x$ *over* $[-1,1]$ *with weight function* $w(x) \equiv 1$.

SOLUTION: We already know a set of orthogonal polynomials ϕ_0, ϕ_1, ϕ_2 on this interval from Example 8; thus we seek $p_2 \in \mathcal{P}_2$ in the form

$$p_2(x) = \gamma_0\phi_0(x) + \gamma_1\phi_1(x) + \gamma_2\phi_2(x). \tag{31}$$

Requiring that

$$f(x) - p_2(x) = e^x - (\gamma_0\phi_0(x) + \gamma_1\phi_1(x) + \gamma_2\phi_2(x))$$

be orthogonal to ϕ_0, ϕ_1 and ϕ_2, i.e. that

$$\int_{-1}^{1} [e^x - (\gamma_0\phi_0(x) + \gamma_1\phi_1(x) + \gamma_2\phi_2(x))]\phi_j(x)\,\mathrm{d}x = 0, \quad j = 0,1,2,$$

we obtain the coefficients γ_j, $j = 0,1,2$:

$$\begin{aligned}
\gamma_0 &= \frac{\int_{-1}^{1} e^x\,\mathrm{d}x}{2} = \frac{e - 1/e}{2}, \\
\gamma_1 &= \frac{\int_{-1}^{1} e^x x\,\mathrm{d}x}{2/3} = \frac{3}{e}, \\
\gamma_2 &= \frac{\int_{-1}^{1} e^x (x - \frac{1}{3})^2\,\mathrm{d}x}{8/45} = \frac{45}{8}\left(\frac{2e}{3} - \frac{14}{3e}\right).
\end{aligned}$$

Substituting the values of γ_0, γ_1 and γ_2 into (31) and recalling the expressions for $\phi_0(x)$, $\phi_1(x)$ and $\phi_2(x)$ from Example 8, we obtain the polynomial of best approximation of degree 2 for the function f. $\diamond$

We conclude by stating a property of orthogonal polynomials that will be required in the next section.

Theorem 9 *Suppose that* $\phi_j(x)$, $j = 0,1,...,$ *is a system of orthogonal polynomials on the interval* (a,b) *with respect to the positive and continuous weight function* $w(x)$. *(It is understood that* $\phi_j(x)$ *is a polynomial of exact degree* j*.) Then, for* $j \geq 1$, *the zeros of the polynomial* $\phi_j(x)$ *are real and distinct, and lie in the interval* (a,b).

PROOF: Suppose that ξ_i, $i = 1, ..., k$, are the points in the open interval (a, b) at which $\phi_j(x)$ changes sign. Let us note that $k \geq 1$, because for $j \geq 1$, by orthogonality of $\phi_j(x)$ to $\phi_0(x) \equiv 1$, we have that

$$\int_a^b w(x)\phi_j(x)\,\mathrm{d}x = 0.$$

Thus the integrand, being a continuous function that is not identically zero on (a, b), must change sign on (a, b); however w is positive on (a, b), so ϕ_j must change sign at least once on (a, b). Therefore $k \geq 1$.

Let us define

$$\pi_k(x) = (x - \xi_1)...(x - \xi_k).$$

Now the function $\phi_j(x)\pi_k(x)$ does not change sign in the interval (a, b), since at each point where $\phi_j(x)$ changes sign $\pi_k(x)$ changes sign also. Hence,

$$\int_a^b w(x)\phi_j(x)\pi_k(x)\,\mathrm{d}x \neq 0.$$

But ϕ_j is orthogonal to every polynomial of lower degree with respect to the weight function w, so the degree of the polynomial π_k must be at least j; thus $k \geq j$. However, k cannot be greater than j, since a polynomial of degree j cannot change sign more than j times. Therefore $k = j$; i.e. the points $\xi_i \in (a, b)$, $i = 1, ..., j$, are the zeros (and all the zeros) of $\phi_j(x)$. □

4 Numerical Integration - Part II

In Section 2 we described the Newton-Cotes family of formulae for numerical integration. These were constructed by replacing the integrand by its Lagrange interpolation polynomial with equally spaced interpolation points and integrating this exactly. Here, we consider another family of numerical integration rules, called Gauss quadrature formulae; these are based on replacing the integrand f by its Hermite interpolation polynomial and choosing the interpolation points x_j in such a way that, upon integrating the Hermite polynomial, the derivative values $f'(x_j)$ do not enter the quadrature formula; it turns out that this can be achieved by requiring that the x_j are roots of a polynomial of a certain degree from a system of orthogonal polynomials.

4.1 Construction of Gauss quadrature rules

Suppose that the function f is defined on the closed interval $[a, b]$ and that it is continuous and differentiable on this interval. Suppose, further, that w is a weight function defined and continuous on $[a, b]$ and positive on (a, b). We wish to construct quadrature formulae for the approximate evaluation of the integral

$$\int_a^b w(x)f(x)\,\mathrm{d}x.$$

Given a non-negative integer n, let x_i, $i = 0, ..., n$, be $(n+1)$ points in the interval $[a, b]$; the precise location of these points will be determined later on. The Hermite interpolation polynomial of degree $(2n+1)$ for the function f is given by the expression (see Section 1.2):

$$p_{2n+1}(x) = \sum_{k=0}^{n} H_k(x) f(x_k) + \sum_{k=0}^{n} K_k(x) f'(x_k),$$

where

$$\begin{aligned} H_k(x) &= [L_k(x)]^2 (1 - 2L_k'(x_k)(x - x_k)), \\ K_k(x) &= [L_k(x)]^2 (x - x_k), \end{aligned}$$

and

$$L_k(x) = \prod_{i=0, i \neq k}^{n} \frac{x - x_i}{x_k - x_i}.$$

Thus,

$$\begin{aligned} \int_a^b w(x) f(x)\, \mathrm{d}x &\approx \int_a^b w(x) p_{2n+1}(x)\, \mathrm{d}x \\ &= \sum_{k=0}^{n} W_k f(x_k) + \sum_{k=0}^{n} V_k f'(x_k), \end{aligned}$$

where

$$W_k = \int_a^b w(x) H_k(x)\, \mathrm{d}x, \qquad V_k = \int_a^b w(x) K_k(x)\, \mathrm{d}x.$$

There is an obvious advantage in choosing the points x_k in such a way that all the coefficients V_k are zero, for then the derivative values $f'(x_k)$ would not be required. Recalling the form of the polynomial $K_k(x)$ and inserting it into the defining expression for V_k, we have

$$\begin{aligned} V_k &= \int_a^b w(x) [L_k(x)]^2 (x - x_k)\, \mathrm{d}x \\ &= \left(\prod_{i=0, i \neq k}^{n} (x_k - x_i)^{-1} \right) \int_a^b w(x) \pi_{n+1}(x) L_k(x)\, \mathrm{d}x, \end{aligned}$$

where $\pi_{n+1}(x) = (x - x_0)...(x - x_n)$. Since π_{n+1} is of degree $(n+1)$ while $L_k(x)$ is of degree n for each k, $0 \leq k \leq n$, each V_k will be zero if the polynomial π_{n+1} is orthogonal to every polynomial of lower degree. We can therefore construct the required quadrature formula by choosing the points x_k, $k = 0, ..., n$, to be the zeros of the polynomial of degree $(n+1)$ in the sequence of orthogonal polynomials over the interval (a, b) with respect to the weight function w; we know from Theorem 9 that these zeros are real and distinct, and all lie in the open interval (a, b).

Having chosen the location of the points x_k, we now consider W_k:

$$\begin{aligned} W_k &= \int_a^b w(x)H_k(x)\,\mathrm{d}x \\ &= \int_a^b w(x)[L_k(x)]^2(1-2L_k'(x_k)(x-x_k))\,\mathrm{d}x \\ &= \int_a^b w(x)[L_k(x)]^2\,\mathrm{d}x - 2L_k'(x_k)V_k. \end{aligned}$$

Since $V_k = 0$, the second term in the last line vanishes and thus we obtain the following numerical integration formula, known as **Gauss quadrature**[18] rule:

$$\int_a^b w(x)f(x)\,\mathrm{d}x \approx \sum_{k=0}^{n} W_k f(x_k), \tag{32}$$

where the **quadrature weights** are

$$W_k = \int_a^b w(x)[L_k(x)]^2\,\mathrm{d}x, \tag{33}$$

and the **quadrature points** x_k, $k = 0, ..., n$, are chosen as the zeros of the polynomial of degree $(n+1)$ from a system of orthogonal polynomials over the interval (a, b) with respect to the weight function w.

Exercise 10 *Find a non-negative integer n, as large as possible, and real numbers A_1, A_2, x_1 and x_2 such that the quadrature rule*

$$\int_{-1}^{1} f(x)\,dx \approx A_1 f(x_1) + A_2 f(x_2) \tag{34}$$

is exact for all $f \in \mathcal{P}_{2n+1}$.

SOLUTION: We have to determine four unknowns A_1, A_2, x_1 and x_2, so we need four equations; thus we take, in turn, $f(x) \equiv 1$, $f(x) = x$, $f(x) = x^2$ and $f(x) = x^3$ and demand that the quadrature rule (34) is exact (namely, the integral of f is equal to the corresponding approximation obtained by inserting f into the right-hand side of (34)). Hence,

$$\begin{aligned} 2 &= A_1 + A_2, && (35)\\ 0 &= A_1x_1 + A_2x_2, && (36)\\ \frac{2}{3} &= A_1x_1^2 + A_2x_2^2, && (37)\\ 0 &= A_1x_1^3 + A_2x_2^3. && (38) \end{aligned}$$

It remains to solve this system. To do so, we consider the quadratic polynomial

$$\pi_2(x) = (x - x_1)(x - x_2)$$

[18] Carl Friedrich Gauss (1777–1855)

whose roots are the unknown quadrature points x_1 and x_2. In expanded form, $\pi_2(x)$ can be written as

$$\pi_2(x) = x^2 + px + q.$$

First we shall determine p and q; then we shall find the roots x_1 and x_2 of π_2. We shall than insert the values of x_1 and x_2 into (36) and solve the linear system (35), (36) for A_1 and A_2.

To find p and q, we multiply (35) by q, (36) by p and (37) by 1, and we add up the resulting equations to deduce that

$$\begin{aligned} \frac{2}{3} + 2q &= A_1(x_1^2 + px_1 + q) + A_2(x_2^2 + px_2 + q) \\ &= A_1\pi_2(x_1) + A_2\pi_2(x_2) = A_1 \cdot 0 + A_2 \cdot 0 = 0. \end{aligned}$$

Therefore,

$$\frac{2}{3} + 2q = 0 \qquad \text{i.e.} \qquad q = -\frac{1}{3}.$$

Similarly, we multiply (36) by q, (37) by p and (38) by 1, and we add up the resulting equations to obtain

$$\begin{aligned} \frac{2}{3}p &= A_1x_1(x_1^2 + px_1 + q) + A_2x_2(x_2^2 + px_2 + q) \\ &= A_1x_1\pi_2(x_1) + A_2x_2\pi_2(x_2) = A_1 \cdot 0 + A_2 \cdot 0 = 0. \end{aligned}$$

Thus,

$$\frac{2}{3}p = 0 \qquad \text{i.e.} \qquad p = 0.$$

Having determined p and q, we see that

$$\pi_2(x) = x^2 - \frac{1}{3},$$

so that

$$x_1 = -\frac{1}{\sqrt{3}}, \qquad x_2 = \frac{1}{\sqrt{3}}.$$

With these values of x_1 and x_2 we find from (35) and (36) that

$$\begin{aligned} A_1 + A_2 &= 2 \\ A_1 - A_2 &= 0, \end{aligned}$$

and therefore $A_1 = A_2 = 1$. To summarise, the required quadrature rule is

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx f(-\frac{1}{\sqrt{3}}) + f(\frac{1}{\sqrt{3}});$$

it is exact for any f in $\mathcal{P}_3$ (i.e. $n = 1$). A straightforward calculation shows that, in general, this quadrature rule is not exact for polynomials of degree higher than 3 (take $f(x) = x^4$, for example). Finally, we note that x_1 and x_2 are the zeros of an orthogonal polynomial, as predicted by the theory; indeed, we see from Example 8 that $x^2 - \frac{1}{3}$ is the Legendre polynomial of degree 2. $\diamond$

4.2 Error estimation for Gauss quadrature

The next theorem provides a bound on the error that has been committed by approximating the integral on the left-hand side of (32) by the quadrature rule on the right.

Theorem 10 *Suppose that w is a weight function, defined and continuous on the closed interval $[a,b]$ and positive on (a,b), and that f is defined and continuous on $[a,b]$; suppose further that f has a continuous derivative of order $(2n+2)$ on $[a,b]$. Then, there exists a number η in (a,b) such that*

$$\int_a^b w(x)f(x)\,dx - \sum_{k=0}^{n} W_k f(x_k) = \frac{f^{(2n+2)}(\eta)}{(2n+2)!}\int_a^b w(x)[\pi_{n+1}(x)]^2\,dx.$$

Consequently, the integration formula (32), (33) will give the exact result for every polynomial of degree $(2n+1)$.

Proof: Recalling Theorem 2 and the definition of the Hermite interpolation polynomial $p_{2n+1}(x)$ for the function f,

$$\begin{aligned}\int_a^b w(x)f(x)\,\mathrm{d}x - \sum_{k=0}^{n} W_k f(x_k) &= \int_a^b w(x)(f(x)-p_{2n+1}(x))\,\mathrm{d}x \\ &= \int_a^b w(x)\frac{f^{(2n+2)}(\xi(x))}{(2n+2)!}[\pi_{n+1}(x)]^2\,\mathrm{d}x.\end{aligned}$$

However, by the Integral Mean Value Theorem[19], the last term is equal to

$$\frac{f^{(2n+2)}(\eta)}{(2n+2)!}\int_a^b w(x)[\pi_{n+1}(x)]^2\,\mathrm{d}x,$$

for some $\eta \in (a,b)$, and hence the desired error estimate. □

Note that, by virtue of Theorem 10, the Gauss quadrature rule gives the exact value of the integral when f is a polynomial of degree $(2n+1)$ or less, which is the highest possible degree that one can hope for with the $(2n+2)$ free parameters consisting of the quadrature weights W_k, $k=0,...,n$, and the quadrature points x_k, $k=0,...,n$.

[19] **Integral Mean Value Theorem:** Suppose that the functions g and h are defined and continuous on the closed interval $[a,b]$, and that h does not change sign in this interval. Then there exists a number η in (a,b) such that

$$\int_a^b g(x)h(x)\,\mathrm{d}x = g(\eta)\int_a^b h(x)\,\mathrm{d}x.$$

5 Piecewise Polynomial Approximation

So far we have concentrated on approximating a given function f defined on an interval $[a,b]$ by a polynomial on that interval either through (Lagrange or Hermite) interpolation, or by seeking the polynomial of best approximation (in the L^∞ or L^2 norm). Each of these constructions is global in nature, in the sense that the approximation is defined by the same analytical expression on the whole interval $[a,b]$; furthermore, if the definition of the function is altered locally (at an interpolation point in the case of Lagrange and Hermite interpolation, or on a subinterval of $[a,b]$ in the case of best approximation in the L^∞ or L^2 norm), the definition of the approximating polynomial changes globally, on the whole interval, and it has to be reconstructed from scratch. An alternative, and more flexible, way of approximating a function f is to divide the interval $[a,b]$ into a number of subintervals and to look for a piecewise approximation by polynomials of low degree. Such piecewise-polynomial approximations are called **splines**. To give a flavour of the theory of splines we consider some simple examples: linear splines and cubic splines.

5.1 Linear splines

Definition 4 *Suppose that f is a real-valued function, defined and continuous on the closed interval $[a,b]$. Further, let $K = \{z_0, ..., z_m\}$ be a subset of $[a,b]$, with $a = z_0 < z_1 < ... < z_{m-1} < z_m = b$. The* **linear spline** *$s_L(x)$, interpolating f at the points z_i, is defined by*

$$s_L(x) = \frac{z_i - x}{z_i - z_{i-1}} f(z_{i-1}) + \frac{x - z_{i-1}}{z_i - z_{i-1}} f(z_i), \quad x \in [z_{i-1}, z_i], \quad i = 1, ..., m.$$

The points z_i, $i = 0, ..., m$, are called the **knots** *of the spline, and K is referred to as the* **set of knots**.

Clearly $s_L(z_i) = f(z_i)$, $i = 0, ..., m$, so s_L is a piecewise linear function on the interval $[a,b]$ which coincides with the function f at the knots.

Given a set of knots $K = \{z_0, ..., z_m\}$, we shall use the notation $h_i = z_i - z_{i-1}$, and let $h = \max_i h_i$. Also, for a positive integer n, we denote by $C^n[a,b]$ the set of all real-valued functions, defined and continuous on the closed interval $[a,b]$, such that all derivatives, up to and including order n, are defined and continuous on $[a,b]$.

In order to highlight the accuracy of interpolation by linear splines we state the following error bound in the L^∞ norm.

Theorem 11 *Suppose that $f \in C^2[a,b]$ and let s_L be the linear spline that interpolates f at the knots $a = z_0 < z_1 < ... < z_{m-1} < z_m = b$; then the following error bound holds:*

$$\|f - s_L\|_\infty \leq \frac{1}{8} h^2 \|f''\|_\infty,$$

where $h = \max_i h_i = \max_i (z_i - z_{i-1})$, and $\|\cdot\|_\infty$ denotes the L^∞ norm on the interval $[a,b]$ defined by $\|u\|_\infty := \max_{x \in [a,b]} |u(x)|$.

PROOF: Consider a subinterval $[z_{i-1}, z_i]$, $1 \le i \le m$. According to Theorem 1, applied on the interval $[z_{i-1}, z_i]$,

$$f(x) - s_L(x) = \frac{1}{2} f''(\xi)(x - z_{i-1})(x - z_i), \qquad x \in [z_{i-1}, z_i],$$

where $\xi = \xi(x) \in (z_{i-1}, z_i)$. Thence

$$\begin{aligned} |f(x) - s_L(x)| &\le \frac{1}{8} h_i^2 \max_{\zeta \in [z_{i-1}, z_i]} |f''(\zeta)|. \\ &\le \frac{1}{8} h^2 \|f''\|_\infty, \end{aligned}$$

for each $x \in [z_{i-1}, z_i]$ and each $i = 1, ..., m$. Hence the required error bound. □

If the interpolation error is measured in the L^2 norm instead, we have an analogous result. In order to state this, we define the function space $H^1(a, b)$ as the set of all $v \in L^2(a, b)$ that are differentiable everywhere on (a, b) except, perhaps, on a set of measure zero, and $v' \in L^2(a, b)$. Let us note, for example, that a linear spline s_L belongs to $H^1(a, b)$. Similarly, we define the function space $H^2(a, b)$ as the set of all v in $H^1(a, b)$ such that v' is differentiable everywhere on (a, b), except perhaps on a set of measure zero, and $v'' \in L^2(a, b)$. The function spaces $H^1(a, b)$ and $H^2(a, b)$ are called **Sobolev spaces.**

Theorem 12 *Suppose that $f \in H^2(a, b)$ and let s_L be the linear spline that interpolates f at the knots $a = z_0 < z_1 < ... < z_{m-1} < z_m = b$; then the following error bounds hold:*

$$\begin{aligned} \|f - s_L\|_2 &\le \left(\frac{h}{\pi}\right)^2 \|f''\|_2, \\ \|f' - s_L'\|_2 &\le \frac{h}{\pi} \|f''\|_2, \end{aligned}$$

where $h = \max_i h_i = \max_i (z_i - z_{i-1})$, and $\|\cdot\|_2$ denotes the L^2 norm on the interval $[a, b]$ defined by $\|u\|_2 := \left(\int_a^b |u(x)|^2 \, dx\right)^{1/2}$.

PROOF: Consider a subinterval $[z_{i-1}, z_i]$, $1 \le i \le m$, and define $e(x) = f(x) - s_L(x)$ for $x \in [z_{i-1}, z_i]$. Then $e \in H^2(z_{i-1}, z_i)$ and $e(z_i) = e(z_{i+1}) = 0$. Therefore e can be expanded into a convergent Fourier sine-series,

$$e(x) = \sum_{k=1}^{\infty} a_k \sin \frac{k\pi(x - z_{i-1})}{h_i}.$$

Hence,

$$\int_{z_{i-1}}^{z_i} [e(x)]^2 \, \mathrm{d}x = \frac{h_i}{2} \sum_{k=1}^{\infty} |a_k|^2.$$

Differentiating the Fourier sine-series for e twice, we deduce that the Fourier coefficients of e' are $(k\pi/h_i)a_k$, while those of e'' are $-(k\pi/h_i)^2 a_k$. Thus,

$$\begin{aligned}\int_{z_{i-1}}^{z_i} [e'(x)]^2 \,\mathrm{d}x &= \frac{h_i}{2} \sum_{k=1}^{\infty} \left(\frac{k\pi}{h_i}\right)^2 |a_k|^2, \\ \int_{z_{i-1}}^{z_i} [e''(x)]^2 \,\mathrm{d}x &= \frac{h_i}{2} \sum_{k=1}^{\infty} \left(\frac{k\pi}{h_i}\right)^4 |a_k|^2.\end{aligned}$$

Because $k^4 \geq k^2 \geq 1$, it follows that

$$\begin{aligned}\int_{z_{i-1}}^{z_i} [e(x)]^2 \,\mathrm{d}x &\leq \left(\frac{h_i}{\pi}\right)^4 \int_{z_{i-1}}^{z_i} [e''(x)]^2 \,\mathrm{d}x, \\ \int_{z_{i-1}}^{z_i} [e'(x)]^2 \,\mathrm{d}x &\leq \left(\frac{h_i}{\pi}\right)^2 \int_{z_{i-1}}^{z_i} [e''(x)]^2 \,\mathrm{d}x.\end{aligned}$$

However $e''(x) = f''(x) - s_L''(x) = f''(x)$ for $x \in (z_{i-1}, z_i)$ because s_L is a linear function on this interval. Therefore, upon summation over $i = 1, ..., m$, and letting $h = \max_i h_i$, we obtain

$$\begin{aligned}\|e\|_2^2 &\leq \left(\frac{h}{\pi}\right)^4 \|f''\|_2^2, \\ \|e'\|_2^2 &\leq \left(\frac{h}{\pi}\right)^2 \|f''\|_2^2.\end{aligned}$$

Upon taking the square root and recalling that $e = f - s_L$ these yield the desired bounds on the interpolation error. □

We conclude this section with a result that provides a characterisation of linear splines from the point of view of Calculus of Variations.

Theorem 13 *Suppose that s_L is a linear spline that interpolates $f \in C[a,b]$ at the knots $a = z_0 < z_1 < ... < z_{m-1} < z_m = b$. Then, for each function v in $H^1(a,b)$ which also interpolates f at these knots,*

$$\|s_L'\|_2 \leq \|v'\|_2.$$

Proof: Let us observe that

$$\begin{aligned}\|v'\|_2^2 &= \int_a^b (v'(x) - s_L'(x))^2 \,\mathrm{d}x + \int_a^b |s_L'(x)|^2 \,\mathrm{d}x \\ &\quad + 2\int_a^b (v'(x) - s_L'(x)) s_L'(x) \,\mathrm{d}x.\end{aligned}$$

We shall now use integration by parts to show that the last term is equal to zero; the desired inequality will then follow by noting that the first term on the right-hand

side is non-negative. Clearly,

$$\int_a^b (v'(x) - s_L'(x))s_L'(x)\,\mathrm{d}x = \sum_{k=1}^{m} \int_{z_{k-1}}^{z_k} (v'(x) - s_L'(x))s_L'(x)\,\mathrm{d}x$$
$$= \sum_{k=1}^{m} [(v(z_k) - s_L(z_k))s_L'(z_k-) - (v(z_{k-1}) - s_L(z_{k-1}))s_L'(z_{k-1}+)$$
$$- \int_{z_{k-1}}^{z_k} (v(x) - s_L(x))s_L''(x)\,\mathrm{d}x]. \qquad (39)$$

Now $v(z_i) - s_L(z_i) = f(z_i) - f(z_i) = 0$ for $i = 0, ..., m$ and, since s_L is a linear polynomial on each interval (z_{k-1}, z_k), $k = 1, ..., m$, it follows that s_L'' is identically zero on each of these intervals. Thus the expression in the square bracket in (39) is equal to zero for each $k = 1, ..., m$. □

Remark 2 *We note that, instead of looking at each section of the interval between two knots, we can express s_L in closed form as*

$$s_L(x) = \sum_{k=0}^{m} \Phi_k(x) f(x_k),$$

where

$$\Phi_k(x) = \begin{cases} 0 & \text{if } x \le z_{k-1} \\ (x - z_{k-1})/h_k & \text{if } z_{k-1} \le x \le z_k \\ (z_{k+1} - x)/h_{k+1} & \text{if } z_k \le x \le z_{k+1} \\ 0 & \text{if } z_{k+1} \le x, \end{cases}$$

for $k = 1, ..., m-1$, with Φ_0 and Φ_m being defined analogously[20]*. This follows by observing that Φ_k is a piecewise linear function on the interval $[a, b]$ with*

$$\Phi_k(x_l) = \begin{cases} 1, & k = l \\ 0, & k \ne l. \end{cases}$$

Thus the linear spline s_L with knots $a = z_0 < z_1 < ... < z_{m-1} < z_m = b$ can be expressed as a linear combination of the 'basis splines' (or, briefly, B-splines) Φ_k. The precise definition of a B-spline will be given in Section 5.4.

5.2 Cubic splines

Suppose that $f \in C[a, b]$ and let $K = \{z_0, ..., z_m\}$ be a set of $(m + 1)$ knots in the interval $[a, b]$, $a = z_0 < z_1 < ... < z_{m-1} < z_m = b$. Consider the set $\mathcal{S}$ of all functions $s \in C^2[a, b]$ such that:

[20] For example,

$$\Phi_0 = \begin{cases} (z_1 - x)/h_1 & \text{if } a = z_0 \le x \le z_1 \\ 0 & \text{if } z_1 \le x. \end{cases}$$

1) $s(z_i) = f(z_i)$, $i = 0, ..., m$,

2) s is a cubic polynomial on $[z_{i-1}, z_i]$, $i = 1, ..., m$.

Any element of $\mathcal{S}$ is called a **cubic spline**. Note that, unlike linear splines which are uniquely determined by the interpolating conditions, there is more than one cubic spline that satisfies the conditions 1) and 2); indeed, there are $4m$ coefficients of cubic polynomials (4 on each of the m subintervals), and only $(m+1)$ interpolating conditions and $3(m-1)$ continuity conditions[21], giving $(4m-2)$ conditions. Hence $\mathcal{S}$ is a linear space of dimension 2.

An important class of cubic splines is singled out by the following definition.

Definition 5 *The* **natural cubic spline**, *denoted s_2 is the element of the set $\mathcal{S}$ satisfying the conditions*

$$s_2''(z_0) = s_2''(z_m) = 0.$$

We shall prove that this definition is correct in the sense that the two additional conditions in Definition 5 uniquely determine s_2: this will be done by describing an algorithm for constructing s_2.

Construction of the natural cubic spline. Let us define $\sigma_i = s_2''(z_i)$, and note that as s_2'' is a linear function on each subinterval $[z_{i-1}, z_i]$; then s_2 can be expressed as

$$s_2''(x) = \frac{z_i - x}{h_i}\sigma_{i-1} + \frac{x - z_{i-1}}{h_i}\sigma_i, \qquad x \in [z_{i-1}, z_i].$$

Integrating this twice we obtain

$$s_2(x) = \frac{(z_i - x)^3}{6h_i}\sigma_{i-1} + \frac{(x - z_{i-1})^3}{6h_i}\sigma_i + \alpha_i(x - z_{i-1}) + \beta_i(z_i - x), \ x \in [z_{i-1}, z_i], \quad (40)$$

where α_i and β_i are constants of integration. Equating s_2 with f at the knots z_{i-1}, z_i yields

$$\begin{aligned} f(z_{i-1}) &= \frac{1}{6}\sigma_{i-1}h_i^2 + h_i\beta_i, \\ f(z_i) &= \frac{1}{6}\sigma_i h_i^2 + h_i\alpha_i. \end{aligned}$$

Determining α_i and β_i from these, inserting them into (40) and exploiting the continuity of s_2' at the internal knots, namely that $s_2'(z_i-) = s_2'(z_i+)$, $i = 1, ..., m-1$, gives

$$h_i\sigma_{i-1} + 2(h_{i+1} + h_i)\sigma_i + h_{i+1}\sigma_{i+1} = 6\left(\frac{f(z_{i+1}) - f(z_i)}{h_{i+1}} - \frac{f(z_i) - f(z_{i-1})}{h_i}\right)$$

[21] Recall that $s \in C^2[a, b]$, so s, s' and s'' are continuous at the internal knots $z_1, ..., z_{m-1}$.

for $i = 1, ..., m-1$, together with

$$\sigma_0 = \sigma_m = 0,$$

which is a system of linear equations for the σ_i; the matrix of the system is tri-diagonal and non-singular. By solving this linear system we obtain the σ_i, $i = 0, ..., m$, and thereby all the α_i, β_i, $i = 1, ..., m$.

We have seen in the previous section, in Theorem 13, that a linear spline can be characterised as a minimiser of a certain quadratic functional. Natural cubic splines have an analogous property.

Theorem 14 *Suppose that s_2 is the natural cubic spline that interpolates a function $f \in H^2(a,b)$ at the knots $a = z_0 < z_1 < ... < z_{m-1} < z_m = b$. Then, for each function v in $H^2(a,b)$ which also interpolates f at the knots,*

$$\|s_2''\|_2 \leq \|v''\|_2.$$

The proof is analogous to that of Theorem 13 and is left as an exercise.

5.3 Hermite cubic splines

In the previous section we took $f \in C[a,b]$; here we shall strengthen our requirements on the smoothness of the function that we wish to interpolate and assume that $f \in C^1[a,b]$. Let $K = \{z_0, ..., z_m\}$ be a set of knots in the interval $[a,b]$, $a = z_0 < z_1 < ... < z_{m-1} < z_m = b$. We define the **Hermite cubic spline** as a function $s \in C^1[a,b]$ such that[22]:

1) $s(z_i) = f(z_i)$, $s'(z_i) = f'(z_i)$, $\quad i = 0, ..., m$,

2) s is a cubic polynomial of $[z_{i-1}, z_i]$, $i = 1, ..., m$.

Writing the spline s on the interval $[z_{i-1}, z_i]$ as

$$s(x) = c_0 + c_1(x - z_{i-1}) + c_2(x - z_{i-1})^2 + c_3(x - z_{i-1})^3, \qquad x \in [z_{i-1}, z_i],$$

we find that $c_0 = f(z_{i-1})$, $c_1 = f'(z_{i-1})$, and

$$\begin{aligned} c_2 &= 3\frac{f(z_i) - f(z_{i-1})}{h_i^2} - \frac{f'(z_i) + 2f'(z_{i-1})}{h_i}, \\ c_3 &= \frac{f'(z_i) + f'(z_{i-1})}{h_i^2} - 2\frac{f(z_i) - f(z_{i-1})}{h_i^3}. \end{aligned}$$

Unlike natural cubic splines, the coefficients of a Hermite cubic spline on each subinterval can be written down explicitly without the need to solve a tri-diagonal system.

Concerning the size of the interpolation error, we have the following result.

[22]Note that, strictly speaking, a Hermite cubic spline is *not* a cubic spline in the sense of the definition from the previous section because it is not necessarily an element of $C^2[a,b]$.

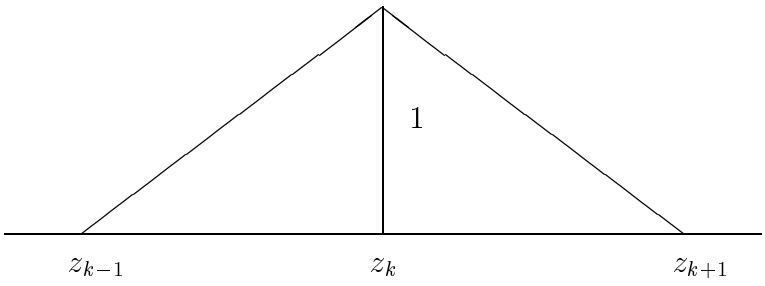

Figure 1: The spline $\Phi_k(x)$.

Theorem 15 *Suppose that $f \in C^4[a,b]$ and let s be the Hermite cubic spline that interpolates f at the knots $a = z_0 < z_1 < ... < z_{m-1} < z_m = b$; then the following error bound holds:*

$$\|f - s\|_\infty \leq \frac{1}{384} h^4 \|f^{(4)}\|_\infty,$$

where $h = \max_i h_i = \max_i (z_i - z_{i-1})$, and $\|\cdot\|_\infty$ denotes the L^∞ norm on the interval $[a,b]$.

The proof is analogous to that of Theorem 11, except that Theorem 2 is used instead of Theorem 1.

5.4 B-splines

In Section 5.1, Remark 2, we saw that any linear spline s_L with knots $a = z_0 < z_1 < ... < z_{m-1} < z_m = b$ can be expressed as a linear combination of 'basis splines', Φ_k, $k = 0, ..., m$. Each Φ_k is a non-negative function that is identically zero outside the interval $[z_{k-1}, z_{k+1}]$, $k = 1, ..., m-1$, as depicted in Figure 1, while Φ_0 and Φ_m are identically zero outside $[a, z_1]$ and $[z_{m-1}, b]$, respectively. In this section we generalise this idea and construct piecewise polynomials of higher degree that have these same properties. In order to proceed, we need the concept of *divided difference*.

Definition 6 *Given that $f \in C[a,b]$, and $a \leq x_0 < x_1 < ... < x_{n-1} < x_n \leq b$, consider the Lagrange interpolation polynomial p_n of degree n with interpolation points x_i, $i = 0, ..., n$, for the function f. The nth* **divided difference** *of f, written $[x_0, ..., x_n]f$, is defined as the coefficient of x^n in the polynomial $p_n(x)$. We define $[x_i]f$ to be just $f(x_i)$.*

From the definition of the Lagrange interpolation polynomial for the function f we deduce that, for $n \geq 1$,

$$[x_0, ..., x_n]f = \sum_{k=0}^{n} \frac{f(x_k)}{\prod_{i=0,\, i\neq k}^{n}(x_k - x_i)}.$$

Before we state the formal definition of a B-spline, let us consider, for $n \geq 1$, the function

$$g_n(z, x) := [(z - x)_+]^{n-1},$$

where, for a real number x, x_+ denotes the positive part of x, i.e. $x_+ = x$ if $x > 0$, and equal to zero otherwise. We take the nth divided difference of g_n with respect to z over a set of $n + 1$ consecutive knots and define:

$$M_{n,i}(x) = [z_{i-n}, ..., z_i]g_n(\cdot, x), \qquad n \leq i \leq m.$$

In particular,

$$M_{1,i}(x) = \begin{cases} (z_i - z_{i-1})^{-1} & \text{for } z_{i-1} \leq x < z_i \\ 0 & \text{otherwise.} \end{cases}$$

Each $M_{n,i}$ is a linear combination of the truncated powers $(z_j - x)_+^{n-1}$ and is therefore a continuous piecewise polynomial function: between two consecutive knots $M_{n,i}(x)$ is a polynomial of degree $n-1$. The function $M_{n,i}$ is called an **unnormalised B-spline**.

Theorem 16 *The unnormalised B-spline, $M_{n,i}(x) = 0$ for x outside the interval $[z_{i-n}, z_i]$.*

In fact, it can be also shown that M_{ni} is a non-negative function, but the proof of this is more technical and will not be given here (see, for example, J. Stoer & R. Bulirsch, *Introduction to Numerical Analysis*, Second Edition, Texts in Applied Mathematics, 12; Springer-Verlag, 1993).

PROOF: For $x < z_{i-n} \leq z \leq z_i$, $g_n(z, x) = (z - x)^{n-1}$ is a polynomial of degree $n - 1$ in z, and therefore it has a vanishing nth divided difference:

$$[z_{i-n}, ..., z_i]g_n(\cdot, x) \equiv 0.$$

Therefore, $M_{n,i}(x) \equiv 0$ for $x < z_{i-n}$. On the other hand, if $z_{i-n} \leq z \leq z_i \leq x$, then $g_n(z, x) = [(z - x)_+]^{n-1} \equiv 0$ is trivially true, so that again $M_{i,n}(x) \equiv 0$. □

Sometimes it is convenient to rescale the unnormalised B-spline so that its values lie in the interval $[0, 1]$; this leads to the so-called **normalised B-spline**,

$$N_{n,i}(x) := (z_i - z_{i-n})M_{n,i}(x).$$

B-splines are simple and easy to manipulate, which makes them particularly attractive in application areas such as computer aided design and computer graphics where fast and stable interpolation processes are of fundamental importance.

6 Approximation of Initial Value Problems for ODEs

Ordinary differential equations frequently occur in mathematical models that arise in many branches of science, engineering and economy. Unfortunately it is seldom that these equations have solutions that can be expressed in closed form, so it is common to seek approximate solutions by means of numerical methods; nowadays this can usually be achieved very inexpensively to high accuracy and with a reliable bound on the error between the analytical solution and its numerical approximation. In this section we shall be concerned with the construction and the analysis of numerical methods for first-order differential equations of the form

$$y' = f(x, y) \tag{41}$$

for the real-valued function y of the real variable x, where $y' \equiv \frac{dy}{dx}$. In order to select a particular integral from the infinite family of solution curves that constitute the general solution to (41), the differential equation will be considered in tandem with an **initial condition**: given two real numbers x_0 and y_0, we seek a solution to (41) for $x > x_0$ such that

$$y(x_0) = y_0. \tag{42}$$

The differential equation (41) together with the initial condition (42) is called an **initial value problem**.

In general, even if $f(\cdot, \cdot)$ is a continuous function, there is no guarantee that the initial value problem (41), (42) possesses a unique solution[23]. Fortunately, under a further mild condition on the function f, the existence and uniqueness of a solution to (41), (42) can be ensured: the result is encapsulated in the next theorem; for a proof, see, for example, P. J. Collins, *Differential and Integral Equations, Part I,* Mathematical Insitute Oxford, 1988 (reprinted 1990).

Theorem 17 (Picard's Theorem[24].) *Suppose that $f(\cdot, \cdot)$ is a continuous function of its arguments in a region U of the (x, y) plane which contains the rectangle*

$$R = \{(x, y) : x_0 \leq x \leq X_M, \quad |y - y_0| \leq Y_M\},$$

where $X_M > x_0$ and $Y_M > 0$ are constants. Suppose also, that there exists a positive constant L such that

$$|f(x, y) - f(x, z)| \leq L|y - z| \tag{43}$$

holds whenever (x, y) and (x, z) lie in the rectangle R. Finally, letting

$$M = \sup\{|f(x, y)| : (x, y) \in R\},$$

[23] Consider, for example, the initial value problem $y' = y^{2/3}$, $y(0) = 0$; this has two solutions: $y(x) \equiv 0$ and $y(x) = x^3/27$.

[24] Emile Picard (1856–1941)

suppose that $M(X_M - x_0) \leq Y_M$. *Then there exists a unique continuously differentiable function* $x \mapsto y(x)$, *defined on the closed interval* $[x_0, X_M]$, *which satisfies (41) and (42).*

The condition (43) is called a **Lipschitz condition**[25], and L is called the **Lipschitz constant** for f.

In the rest of this section we shall consider numerical methods for the approximate solution of the initial value problem (41), (42). We shall suppose throughout that the function f satisfies the conditions of Picard's Theorem on the rectangle R and that the initial value problem has a unique solution defined on the interval $[x_0, X_M]$. We begin by discussing one-step methods; this will be followed by the study of linear multi-step methods.

6.1 One-step methods

The simplest example of a one-step method for the numerical solution of the initial value problem (41), (42) is Euler's method[26].

Euler's method. Suppose that the initial value problem (41), (42) is to be solved on the interval $[x_0, X_M]$. We divide this interval by the **mesh-points** $x_n = x_0 + nh$, $n = 0, ..., N$, where $h = (X_M - x_0)/N$ and N is a positive integer. The positive real number h is called the **step size**. Now let us suppose that, for each n, we seek a numerical approximation y_n to $y(x_n)$, the value of the analytical solution at the mesh point x_n. Given that $y(x_0) = y_0$, let us suppose that we have already calculated y_n, up to some n, $0 \leq n \leq N - 1$; we define

$$y_{n+1} = y_n + hf(x_n, y_n).$$

Thus taking in succession $n = 0, 1, ..., N - 1$, one step at a time, the approximate values y_n at the mesh points x_n can be easily obtained. This numerical method is known as **Euler's method**. A general one-step method may be written in the form:

$$y_{n+1} = y_n + h\Phi(x_n, y_n; h), \qquad n = 0, ..., N - 1, \quad y(x_0) = y_0, \tag{44}$$

where $\Phi(\cdot, \cdot; \cdot)$ is a continuous function of its variables. For example, in the case of Euler's method, $\Phi(x_n, y_n; h) = f(x_n, y_n)$.

In order to assess the accuracy of the numerical method (44), we define the **global error**, e_n, by

$$e_n = y(x_n) - y_n.$$

We also need the concept of **truncation error**, T_n, defined by

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \Phi(x_n, y(x_n); h). \tag{45}$$

The next theorem provides a bound on the magnitude of the global error in terms of the truncation error.

[25] Rudolf Lipschitz (1832–1903)
[26] Leonard Euler (1707–1783)

Theorem 18 *Consider the general one-step method (44) where, in addition to being a continuous function of its arguments, Φ is assumed to satisfy a Lipschitz condition with respect to its second argument; namely, there exists a positive constant L_Φ such that, for $0 \le h \le h_0$ and for the same region R as in Picard's theorem,*

$$|\Phi(x,y;h) - \Phi(x,z;h)| \le L_\Phi |y-z|, \qquad \text{for } (x,y),\ (x,z) \text{ in } R. \tag{46}$$

Then, assuming that $|y_n - y_0| \le Y_M$, it follows that

$$|e_n| \le \mathrm{e}^{L_\Phi(x_n - x_0)}|e_0| + \left[\frac{\mathrm{e}^{L_\Phi(x_n-x_0)} - 1}{L_\Phi}\right] T, \quad n = 0, ..., N, \tag{47}$$

where $T = \max_{0\le n\le N-1} |T_n|$.

PROOF: Subtracting (44) from (45) we obtain:

$$e_{n+1} = e_n + h[\Phi(x_n, y(x_n); h) - \Phi(x_n, y_n; h)] + hT_n.$$

Then, since $(x_n, y(x_n))$ and (x_n, y_n) belong to R, the Lipschitz condition (46) implies that

$$|e_{n+1}| \le |e_n| + hL_\Phi|e_n| + h|T_n|, \qquad n = 0, ..., N-1.$$

That is,

$$|e_{n+1}| \le (1 + hL_\Phi)|e_n| + h|T_n|, \qquad n = 0, ..., N-1.$$

Hence

$$\begin{aligned}
|e_1| &\le (1+hL_\Phi)|e_0| + hT, \\
|e_2| &\le (1+hL_\Phi)^2|e_0| + h[1 + (1+hL_\Phi)]T, \\
|e_3| &\le (1+hL_\Phi)^3|e_0| + h[1 + (1+hL_\Phi) + (1+hL_\Phi)^2]T, \\
&\text{etc.} \\
|e_n| &\le (1+hL_\Phi)^n|e_0| + [(1+hL_\Phi)^n - 1]T/L_\Phi.
\end{aligned}$$

Observing that $1 + hL_\Phi \le \exp(hL_\Phi)$, we obtain (47). □

Let us apply this general result in order to obtain a bound on the global error in Euler's method. The truncation error for Euler's method is given by

$$\begin{aligned}
T_n &= \frac{y(x_{n+1}) - y(x_n)}{h} - f(x_n, y(x_n)) \\
&= \frac{y(x_{n+1}) - y(x_n)}{h} - y'(x_n).
\end{aligned} \tag{48}$$

Assuming that $y \in C^2[x_0, X_M]$ and expanding $y(x_{n+1})$ about the point x_n into a Taylor series with remainder, we have that

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2!}y''(\xi), \qquad x_n < \xi < x_{n+1}.$$

Substituting this expansion into (48) gives

$$T_n = \frac{1}{2} h y''(\xi).$$

Let $M_2 = \max_{\zeta \in [x_0, X_M]} |y''(\zeta)|$. Then $|T_n| \leq T$, $n = 0, ..., N-1$, where $T = \frac{1}{2} h M_2$. Inserting this into (47) and noting that for Euler's method $\Phi(x_n, y_n; h) \equiv f(x_n, y_n)$ and therefore $L_\Phi = L$ where L is the Lipschitz constant for f (cf. (43)), we have that

$$|e_n| \leq \mathrm{e}^{L(x_n - x_0)} |e_0| + \frac{1}{2} M_2 \left[\frac{\mathrm{e}^{L(x_n - x_0)} - 1}{L} \right] h, \quad n = 0, ..., N. \tag{49}$$

Let us highlight the practical relevance of our error analysis by focusing on a particular example.

Example 10 *Consider the initial value problem* $y' = \tan^{-1} y$, $y(0) = y_0$. *We need to find* L *and* M_2. *Here* $f(x, y) = \tan^{-1} y$; *so, by the Mean Value Theorem,*

$$|f(x, y) - f(x, z)| = \left| \frac{\partial f}{\partial y}(x, \eta) \, (y - z) \right|,$$

where η *lies between* y *and* z. *In our case*

$$\left| \frac{\partial f}{\partial y} \right| = |(1 + y^2)^{-1}| \leq 1,$$

and therefore $L = 1$. *To find* M_2 *we need to obtain a bound on* $|y''|$ *(without actually solving the initial value problem!). This is easily achieved by differentiating both sides of the differential equation with respect to* x:

$$y'' = \frac{d}{dx}(\tan^{-1} y) = (1 + y^2)^{-1} \frac{dy}{dx} = (1 + y^2)^{-1} \tan^{-1} y.$$

Therefore $|y''(x)| \leq M_2 = \frac{1}{2}\pi$. *Inserting the values of* L *and* M_2 *into (49),*

$$|e_n| \leq \mathrm{e}^{x_n} |e_0| + \frac{1}{4} \pi \, (\mathrm{e}^{x_n} - 1) \, h, \quad n = 0, ..., N.$$

In particular if we assume that no error has been committed initially (i.e. $e_0 = 0$*), we have that*

$$|e_n| \leq \frac{1}{4} \pi \, (\mathrm{e}^{x_n} - 1) \, h, \quad n = 0, ..., N.$$

Thus, given a tolerance TOL, *specified beforehand, we can ensure that the error between the (unknown) analytical solution and its numerical approximation does not exceed this tolerance by choosing a positive step size* h *such that*

$$h \leq \frac{4}{\pi} (\mathrm{e}^{X_M} - 1)^{-1} \, TOL;$$

For such h we shall have $|y(x_n) - y_n| = |e_n| \leq TOL$ for each $n = 0, ..., N$, as required. Thus, at least in principle, we can calculate the numerical solution to arbitrarily high accuracy by choosing a sufficiently small step size. Unfortunately, this is virtually impossible to achieve in practice because digital computers use finite-precision arithmetic and there will always be small (but not infinitely small) pollution effects due to rounding errors; however, these can also be bounded by performing an analysis similar to the one above where $f(x_n, y_n)$ is replaced by its finite-precision representation.

Returning to the general one-step method (44), we consider the choice of the function Φ. Theorem 18 suggests that if the truncation error 'approaches zero' as $h \to 0$ then the global error 'converges to zero' also (as long as $|e_0| \to 0$ when $h \to 0$). This observation motivates the following definition.

Definition 7 *The numerical method (44) is* **consistent** *with the differential equation (41) if the truncation error, defined by (45) is such that for any $\epsilon > 0$, there exists a positive $h(\epsilon)$ for which $|T_n| < \epsilon$ for $0 < h < h(\epsilon)$ and any pair of points $(x_n, y(x_n))$, $(x_{n+1}, y(x_{n+1}))$ on any solution curve in R.*

For the general one-step method (44) we have assumed that the function $\Phi(\cdot, \cdot; \cdot)$ is continuous; also y' is a continuous function on $[x_0, X_M]$. Therefore, from (45),

$$\lim_{h \to 0} T_n = y'(x_n) - \Phi(x_n, y(x_n); 0).$$

This implies that the one-step method (44) is consistent if and only if

$$\Phi(x, y; 0) \equiv f(x, y). \tag{50}$$

We shall henceforth always assume that this condition holds.

Now we are ready to state a convergence theorem for the general one-step method (44).

Theorem 19 *Suppose that the solution of the initial value problem (41), (42) lies in R as does its approximation generated from (44) when $h \leq h_0$. Suppose also that the function $\Phi(\cdot, \cdot; \cdot)$ is uniformly continuous on $R \times [0, h_0]$ and satisfies the consistency condition (50) and the Lipschitz condition*

$$|\Phi(x, y; h) - \Phi(x, z; h)| \leq L_\Phi |y - z| \qquad \text{on } R \times [0, h_0]. \tag{51}$$

Then, if successive approximation sequences (y_n), generated for $x_n = x_0 + nh$, $n = 1, 2, ..., N$, are obtained from (44) with successively smaller values of h, each less than h_0, we have convergence of the numerical solution to the solution of the initial value problem in the sense that

$$|y(x_n) - y_n| \to 0 \qquad \text{as } h \to 0,\ x_n \to x \in [x_0, X_M].$$

PROOF: Suppose that $h = (X_M - x_0)/N$, where N is a positive integer. We shall assume that N is sufficiently large so that $h \leq h_0$. Since $y(x_0) = y_0$ and therefore $e_0 = 0$, Theorem 18 implies that

$$|y(x_n) - y_n| \leq \left[\frac{\mathrm{e}^{L_\phi(X_M - x_0)} - 1}{L_\phi}\right] \max_{0 \leq m \leq n-1} |T_m|, \qquad n = 1, ..., N. \tag{52}$$

From the consistency condition (50) we have

$$T_n = \left[\frac{y(x_{n+1}) - y(x_n)}{h} - f(x_n, y(x_n))\right] + [\Phi(x_n, y(x_n); 0) - \Phi(x_n, y(x_n); h)].$$

According to the Mean Value Theorem the expression in the first bracket is equal to $y'(\xi) - y'(x_n)$, where $\xi \in [x_n, x_{n+1}]$. Since $y'(\cdot) = f(\cdot, y(\cdot)) = \Phi(\cdot, y(\cdot); 0)$ and $\Phi(\cdot, \cdot; \cdot)$ is uniformly continuous on $R \times [0, h_0]$, it follows that y' is uniformly continuous on $[x_0, X_M]$. Thus, for each $\epsilon > 0$ there exists $h_1(\epsilon)$ such that

$$|y'(\xi) - y'(x_n)| \leq \frac{1}{2}\epsilon \qquad \text{for } h < h_1(\epsilon),\ n = 0, 1, ..., N - 1.$$

Also, by the uniform continuity of Φ with respect to its third argument, there exists $h_2(\epsilon)$ such that

$$|\Phi(x_n, y(x_n); 0) - \Phi(x_n, y(x_n); h)| \leq \frac{1}{2}\epsilon \qquad \text{for } h < h_2(\epsilon),\ n = 0, 1, ..., N - 1.$$

Thus, defining $h(\epsilon) = \min(h_1(\epsilon), h_2(\epsilon))$, we have

$$|T_n| \leq \epsilon \qquad \text{for } h < h(\epsilon),\ n = 0, 1, ..., N - 1.$$

Inserting this into (52) we deduce that $|y(x_n) - y_n| \to 0$ as $h \to 0$; since

$$|y(x) - y_n| \leq |y(x) - y(x_n)| + |y(x_n) - y_n|,$$

and the first term on the right also converges to zero as $h \to 0$ by the uniform continuity of y on the interval $[x_0, X_M]$, the proof is complete. □

We saw earlier that for Euler's method the magnitude of the truncation error T_n is bounded above by a constant multiple of the step size h, that is

$$|T_n| \leq Kh \qquad \text{for } 0 < h \leq h_0,$$

where K is a positive constant, independent of h. However there are other one-step methods (a class of which, called Runge-Kutta methods, will be considered below) for which we can do better. Thus, in order to quantify the asymptotic rate of decay of the truncation error as the step size h converges to zero, we introduce the following definition.

Definition 8 *The numerical method (44) is said to have* **order of accuracy** p*, if* p *is the largest positive integer such that, for any sufficiently smooth solution curve* $(x, y(x))$ *in* R *of the initial value problem (41), (42), there exist constants* K *and* h_0 *such that*

$$|T_n| \leq Kh^p \qquad for\ 0 < h \leq h_0$$

for any pair of points $(x_n, y(x_n))$, $(x_{n+1}, y(x_{n+1}))$ *on the solution curve.*

Runge-Kutta methods. In the sense of this definition, Euler's method is only first-order accurate; nevertheless, it is simple and cheap to implement because, to obtain y_{n+1} from y_n, we only require a single evaluation of the function f, at (x_n, y_n). Runge-Kutta methods aim to achieve higher accuracy by sacrificing the efficiency of Euler's method through re-evaluating $f(\cdot, \cdot)$ at points intermediate between $(x_n, y(x_n))$ and $(x_{n+1}, y(x_{n+1}))$. Consider, for example, the following family of methods:

$$y_{n+1} = y_n + h(ak_1 + bk_2), \tag{53}$$

where

$$k_1 = f(x_n, y_n), \tag{54}$$
$$k_2 = f(x_n + \alpha h, y_n + \beta h k_1), \tag{55}$$

and where the parameters a, b, α and β are to be determined.[27] Clearly (53) – (55) can be rewritten in the form (44) and therefore it is a family of one step methods. By the condition (50), a method from this family will be consistent if and only if

$$a + b = 1.$$

Further conditions on the parameters are obtained by attempting to maximise the order of accuracy of the method. Indeed, expanding the truncation error of (53) – (55) in powers of h, after some algebra we obtain

$$\begin{aligned} T_n &= \frac{1}{2}hy''(x_n) + \frac{1}{6}h^2y'''(x_n) \\ &\quad -bh[\alpha f_x + \beta f_y f] - bh^2\left[\frac{1}{2}\alpha^2 f_{xx} + \alpha\beta f_{xy}f + \frac{1}{2}\beta^2 f_{yy}f^2\right] + O(h^3). \end{aligned}$$

Here we have used the abbreviations $f = f(x_n, y(x_n))$, $f_x = \frac{\partial f}{\partial x}(x_n, y(x_n))$, etc. On noting that $y'' = f_x + f_y f$, it follows that $T_n = O(h^2)$ for any f, provided

$$\alpha b = \beta b = \frac{1}{2},$$

which implies that if $\beta = \alpha$, $b = \frac{1}{2\alpha}$ and $a = 1 - \frac{1}{2\alpha}$ then the method is second-order accurate; while this still leaves one free parameter, α, it is easy to see that

[27] We note in passing that Euler's method is a member of this family of methods, corresponding to $a = 1$ and $b = 0$. However we are now seeking methods that are at least second-order accurate.

no choice of the parameters will make the method generally third-order accurate. There are two well-known examples of second-order Runge-Kutta methods of the form (53)–(55):

a) **The modified Euler method:** In this case we take $\alpha = \frac{1}{2}$ to obtain

$$y_{n+1} = y_n + h\, f\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hf(x_n, y_n)\right);$$

b) **The improved Euler method:** This is arrived at by choosing $\alpha = 1$ which gives

$$y_{n+1} = y_n + \frac{1}{2}h\left[f(x_n, y_n) + f(x_n + h, y_n + hf(x_n, y_n))\right].$$

For these two methods it is easily verified by Taylor series expansion that the truncation error is of the form, respectively,

$$\begin{aligned} T_n &= \frac{1}{6}h^2\left[f_y(f_x + f_y f) + \frac{1}{4}(f_{xx} + 2f_{xy}f + f_{yy}f^2)\right] + O(h^3), \\ T_n &= \frac{1}{6}h^2\left[f_y(f_x + f_y f) - \frac{1}{2}(f_{xx} + 2f_{xy}f + f_{yy}f^2)\right] + O(h^3). \end{aligned}$$

A particularly popular example of a Runge-Kutta method is the fourth-order method:

$$y_{n+1} = y_n + \frac{1}{6}h\left(k_1 + 2k_2 + 2k_3 + k_4\right),$$

where

$$\begin{aligned} k_1 &= f(x_n, y_n) \\ k_2 &= f\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_1\right) \\ k_3 &= f\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_2\right) \\ k_4 &= f(x_n + h, y_n + hk_3). \end{aligned}$$

Here k_2 and k_3 represent approximations to the derivative $y'(\cdot)$ at points on the solution curve, intermediate between $(x_n, y(x_n))$ and $(x_{n+1}, y(x_{n+1}))$, and $\Phi(x_n, y_n; h)$ is a weighted average of the k_i, $i = 1, ..., 4$, the weights corresponding to those of Simpson's rule (to which the fourth-order Runge-Kutta method reduces when $\frac{\partial f}{\partial y} \equiv 0$).

Exercise 11 (Oxford Finals, 1992) *Let α be a non-zero real number and let $x_n = a + nh$, $n = 0, \ldots, N$, be a uniform mesh on the interval $[a, b]$ of step size $h = (b - a)/N$. Consider the explicit one-step method for the numerical solution of the initial value problem $y' = f(x, y)$, $y(a) = y_0$, which determines approximations y_n to the values $y(x_n)$ from the recurrence relation*

$$y_{n+1} = y_n + h(1 - \alpha)f(x_n, y_n) + h\alpha f\left(x_n + \frac{h}{2\alpha}, y_n + \frac{h}{2\alpha}f(x_n, y_n)\right).$$

Show that this method is consistent and that its truncation error, $T_n(h,\alpha)$, can be expressed as

$$T_n(h,\alpha) = \frac{h^2}{8\alpha}\left[\left(\frac{4}{3}\alpha - 1\right) y'''(x_n) + y''(x_n)\frac{\partial f}{\partial y}(x_n, y(x_n))\right] + O(h^3).$$

This numerical method is applied to the initial value problem $y' = -y^p$, $y(0) = 1$, where p is a positive integer. Show that if $p = 1$ then $T_n(h,\alpha) = O(h^2)$ for every non-zero real number α. Show also that if $p \geq 2$ then there exists a non-zero real number α_0 such that $T_n(h,\alpha_0) = O(h^3)$.

SOLUTION: Let us define

$$\Phi(x,y;h) = (1-\alpha)f(x,y) + \alpha f\left(x + \frac{h}{2\alpha}, y + \frac{h}{2\alpha}f(x,y)\right).$$

Then the numerical method can be rewritten as

$$y_{n+1} = y_n + h\Phi(x_n, y_n; h).$$

Since

$$\Phi(x,y;0) = f(x,y),$$

the method is consistent. By definition, the truncation error is

$$T_n(h,\alpha) = \frac{y(x_{n+1}) - y(x_n)}{h} - \Phi(x_n, y(x_n); h).$$

We shall perform a Taylor expansion of $T_n(h,\alpha)$ to show that it can be expressed in the desired form. Indeed,

$$\begin{aligned}
T_n(h,\alpha) &= y'(x_n) + \frac{h}{2}y''(x_n) + \frac{h^2}{6}y'''(x_n) \\
&\quad -(1-\alpha)y'(x_n) - \alpha f(x_n + \frac{h}{2\alpha}, y(x_n) + \frac{h}{2\alpha}y'(x_n)) + O(h^3) \\
&= y'(x_n) + \frac{h}{2}y''(x_n) + \frac{h^2}{6}y'''(x_n) - (1-\alpha)y'(x_n) \\
&\quad -\alpha\left[f(x_n, y(x_n)) + \frac{h}{2\alpha}f_x(x_n, y(x_n)) + \frac{h}{2\alpha}f_y(x_n, y(x_n))y'(x_n)\right] \\
&\quad -\frac{\alpha}{2}\left[\left(\frac{h}{2\alpha}\right)^2 f_{xx}(x_n, y(x_n)) + 2\left(\frac{h}{2\alpha}\right)^2 f_{xy}(x_n, y(x_n))y'(x_n)\right. \\
&\qquad\qquad \left. + \left(\frac{h}{2\alpha}\right)^2 f_{yy}(x_n, y(x_n))[y'(x_n)]^2\right] + O(h^3) \\
&= y'(x_n) - (1-\alpha)y'(x_n) - \alpha y'(x_n) \\
&\quad +\frac{h}{2}y''(x_n) - \frac{h}{2}\left[f_x(x_n, y(x_n)) + f_y(x_n, y(x_n))y'(x_n)\right] \\
&\quad +\frac{h^2}{6}y'''(x_n) - \frac{h^2}{8\alpha}\left[f_{xx}(x_n, y(x_n)) + 2f_{xy}(x_n, y(x_n))y'(x_n)\right.
\end{aligned}$$

$$+ f_{yy}(x_n, y(x_n))[y'(x_n)]^2\Big] + O(h^3)$$
$$= \frac{h^2}{6}y'''(x_n) - \frac{h^2}{8\alpha}\left[y'''(x_n) - y''(x_n)f_y(x_n, y(x_n))\right] + O(h^3)$$
$$= \frac{h^2}{8\alpha}\left[\left(\frac{4}{3}\alpha - 1\right)y'''(x_n) + y''(x_n)\frac{\partial f}{\partial y}(x_n, y(x_n))\right] + O(h^3),$$

as required.

Now let us apply the method to $y' = -y^p$, with $p \geq 1$. If $p = 1$, then $y''' = -y'' = y' = -y$, so that

$$T_n(h, \alpha) = -\frac{h^2}{6}y(x_n) + O(h^3).$$

As $y(x_n) = \mathrm{e}^{-x_n} \neq 0$, it follows that

$$T_n(h, \alpha) = O(h^2)$$

for all (non-zero) α.

Finally, suppose that $p \geq 2$. Then

$$y'' = -py^{p-1}y' = py^{2p-1}$$

and

$$y''' = p(2p-1)y^{2p-2}y' = -p(2p-1)y^{3p-2},$$

and therefore

$$T_n(h, \alpha) = -\frac{h^2}{8\alpha}\left[\left(\frac{4}{3}\alpha - 1\right)p(2p-1) + p^2\right]y^{3p-2}(x_n) + O(h^3).$$

Choosing α such that

$$\left(\frac{4}{3}\alpha - 1\right)p(2p-1) + p^2 = 0,$$

namely

$$\alpha = \alpha_0 = \frac{3p-3}{8p-4},$$

gives

$$T_n(h, \alpha_0) = O(h^3).$$

We note in passing that for $p > 1$ the exact solution of the initial value problem $y' = -y^p$, $y(0) = 1$, is $y(x) = [(p-1)x + 1]^{1/(1-p)}$. $\diamond$

6.2 Linear multi-step methods

While Runge-Kutta methods present an improvement over Euler's method in terms of accuracy, this is achieved by investing additional computational effort; in fact, Runge-Kutta methods require more evaluations of $f(\cdot, \cdot)$ than would seem necessary. For example, the fourth-order method involves four function evaluations per step. For comparison, by considering three consecutive points x_{n-1}, $x_n = x_{n-1} + h$, $x_{n+1} =$

$x_{n-1}+2h$, integrating the differential equation between x_{n-1} and x_{n+1}, and applying Simpson's rule to approximate the resulting integral yields

$$y(x_{n+1}) = y(x_{n-1}) + \int_{x_{n-1}}^{x_{n+1}} f(x, y(x))\,\mathrm{d}x$$
$$\approx y(x_{n-1}) + \frac{1}{3}h\left[f(x_{n-1}, y(x_{n-1})) + 4f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))\right]$$

which leads to the method

$$y_{n+1} = y_{n-1} + \frac{1}{3}h\left[f(x_{n-1}, y_{n-1}) + 4f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right]. \tag{56}$$

In contrast with the one-step methods considered in the previous section where only a single value y_n was required to compute the next approximation y_{n+1}, here we need *two* preceding values, y_n and y_{n-1} to be able to calculate y_{n+1}, and therefore (56) is not a one-step method.

In this section we consider a class of methods of the type (56) for the numerical solution of the initial value problem (41), (42), called **linear multi-step methods**.

Given a sequence of equally spaced mesh points (x_n) with step size h, we consider the general **linear k-step method**

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y_{n+j}), \tag{57}$$

where the coefficients α_0,...,α_k and β_0,...,β_k are real constants. In order to avoid degenerate cases, we shall assume that $\alpha_k \neq 0$ and that α_0 and β_0 are not both equal to zero. If $\beta_k = 0$ then y_{n+k} is obtained explicitly from previous values of y_j and $f(x_j, y_j)$, and the k-step method is then said to be **explicit**. On the other hand, if $\beta_k \neq 0$ then y_{n+k} appears not only on the left-hand side but also on the right, within $f(x_{n+k}, y_{n+k})$; due to this implicit dependence on y_{n+k} the method is then called **implicit**. The numerical method (57) is called *linear* because it involves only linear combinations of the $\{y_n\}$ and the $\{f(x_n, y_n)\}$; for the sake of notational simplicity, henceforth we shall write f_n instead of $f(x_n, y_n)$.

Example 11 *We have already seen an example of a linear 2-step method in (56); here we present further examples of linear multi-step methods.*

a) Euler's method is a trivial case: it is an explicit linear one-step method. The **implicit Euler method**

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1})$$

is an implicit linear one-step method.

b) The **trapezium method**, *given by*

$$y_{n+1} = y_n + \frac{1}{2}h[f_{n+1} + f_n]$$

is also an implicit linear one-step method.

c) The **Adams**[28]**- Bashforth method**

$$y_{n+4} = y_{n+3} + \frac{1}{24}h[55f_{n+3} - 59f_{n+2} + 37f_{n+1} - 9f_n]$$

is an example of an explicit linear four-step method; the **Adams - Moulton method**

$$y_{n+4} = y_{n+3} + \frac{1}{24}h[9f_{n+4} + 19f_{n+3} - 5f_{n+2} - 9f_{n+1}]$$

is an implicit linear four-step method.

There are systematic ways of generating linear multi-step methods, but these constructions will not be discussed here. Instead, we turn our attention to the analysis of linear multi-step methods and introduce the concepts of stability, consistency and convergence.

6.2.1 Zero stability

As is clear from (57) we need k starting values, $y_0, \ldots, y_{k-1}$, before we can apply a linear k-step method to the initial value problem (41), (42): of these, y_0 is given by the initial condition (42), but the others, $y_1, \ldots, y_{k-1}$, have to be computed by other means: say, by using a suitable Runge-Kutta method. At any rate, the starting values will contain numerical errors and it is important to know how these will affect further approximations y_n, $n \geq k$, which are calculated by means of (57). Thus, we wish to consider the 'stability' of the numerical method with respect to 'small perturbations' in the starting conditions.

Definition 9 *A linear k-step method (for the ordinary differential equation $y' = f(x, y)$) is said to be* **zero-stable** *if there exists a constant K such that, for any two sequences (y_n) and $(\hat{y}_n)$ that have been generated by the same formulae but different initial data y_0, $y_1, ..., y_{k-1}$ and $\hat{y}_0$, $\hat{y}_1, ..., \hat{y}_{k-1}$, respectively, we have*

$$|y_n - \hat{y}_n| \leq K \max\{|y_0 - \hat{y}_0|, |y_1 - \hat{y}_1|, ..., |y_{k-1} - \hat{y}_{k-1}|\} \tag{58}$$

for $x_n \leq X_M$, and as h tends to 0.

We shall prove later on that whether or not a method is zero stable can be determined by merely considering its behaviour when applied to the trivial differential equation $y' = 0$, corresponding to (41) with $f(x, y) \equiv 0$; it is for this reason that the kind of stability expressed in Definition 9 is called *zero* stability. While Definition 9 is expressive in the sense that it conforms with the intuitive notion of stability whereby "small perturbations at input give rise to small perturbations at output", it would be a very tedious exercise to verify the zero-stability of a linear multi-step method using Definition 9 only; thus we shall next formulate an algebraic equivalent

[28] J. C. Adams (1819–1892)

of zero stability, known as the root condition, which will simplify this task. Before doing so we introduce some notation.

Given the linear k-step method (57) we consider its **first** and **second characteristic polynomial**, respectively

$$\begin{aligned} \rho(z) &= \sum_{j=0}^{k} \alpha_j z^j, \\ \sigma(z) &= \sum_{j=0}^{k} \beta_j z^j, \end{aligned}$$

where, as before, we assume that

$$\alpha_k \neq 0, \quad \alpha_0^2 + \beta_0^2 \neq 0.$$

Now we are ready to state the main result of this section.

Theorem 20 *A linear multi-step method is zero stable for any ordinary differential equation of the form (41) where f satisfies the Lipschitz condition (43), if and only if its first characteristic polynomial has zeros inside the closed unit disc, with any which lie on the unit circle being simple.*

The algebraic stability condition contained in this theorem, namely that the roots of the first characteristic polynomial lie in the closed unit disc and those on the unit circle are simple, is often called the **root condition.**

PROOF: *Necessity.* Consider the linear k-step method, applied to $y' = 0$:

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + ... + \alpha_1 y_{n+1} + \alpha_0 y_n = 0. \tag{59}$$

The general solution of this kth order linear difference equation has the form

$$y_n = \sum_s p_s(n) z_s^n, \tag{60}$$

where z_s is a zero of the first characteristic polynomial $\rho(z)$ and the polynomial $p_s(\cdot)$ has degree one less than the multiplicity of the zero. Clearly, if $|z_s| > 1$ then there are starting values for which the corresponding solutions grow like $|z_s|^n$ and if $|z_s| = 1$ and its multiplicity is $m_s > 1$ then there are solutions growing like $n^{m_s - 1}$. In either case there are solutions that grow unboundedly as $n \to \infty$, i.e. as $h \to 0$ with nh fixed. Considering starting data $y_0, y_1, ..., y_{k-1}$ which give rise to such an unbounded solution (y_n), and starting data $\hat{y}_0 = \hat{y}_1 = ... = \hat{y}_{k-1} = 0$ for which the corresponding solution of (59) is $(\hat{y}_n)$ with $\hat{y}_n = 0$ for all n, we see that (58) cannot hold. To summarise, if the root condition is violated then the method is not zero stable.

Sufficiency. The proof that the root condition is sufficient for zero stability is long and technical, and will be omitted here. For details, see, for example, K.W. Morton, *Numerical Solution of Ordinary Differential Equations*, Oxford University Computing Laboratory, 1987, or P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962. □

Example 12 *We shall consider the methods from Example 11.*

a) The explicit and implicit Euler methods have first characteristic polynomial $\rho(z) = z - 1$ *with simple root* $z = 1$*, so both methods are zero stable. The same is true of the trapezium method.*

b) The Adams-Bashforth and Adams-Moulton methods considered in Example 11 have the same first characteristic polynomial, $\rho(z) = z^3(z-1)$*, and therefore both methods are zero stable.*

c) The three-step (sixth order accurate) linear multi-step method

$$11y_{n+3} + 27y_{n+2} - 27y_{n+1} - 11y_n = 3h[f_{n+3} + 9f_{n+2} + 9f_{n+1} + f_n]$$

is not *zero-stable. Indeed, the associated first characteristic polynomial* $\rho(z) = 11z^3 + 27z^2 - 27z - 11$ *has roots at* $z_1 = 1$*,* $z_2 \approx -0.3189$*,* $z_3 \approx -3.1356$*, so* $|z_3| > 1$*.*

6.2.2 Consistency

In this section we consider the accuracy of the linear k-step method (57). For this purpose, as in the case of one-step methods, we introduce the notion of truncation error. Thus, suppose that $y(x)$ is a solution of the ordinary differential equation (41). Then the truncation error of (57) is defined as follows:

$$T_n = \frac{\sum_{j=0}^{k} [\alpha_j y(x_{n+j}) - h\beta_j y'(x_{n+j})]}{h\sum_{j=0}^{k} \beta_j}. \tag{61}$$

Of course, the definition requires implicitly that $\sigma(1) = \sum_{j=0}^{k} \beta_j \neq 0$. Again, as in the case of one-step methods, the truncation error can be thought of as the residual that is obtained by inserting the solution of the differential equation into the formula (57) and scaling this residual appropriately (in this case dividing through by $h\sum_{j=0}^{k} \beta_j$), so that T_n resembles $y' - f(x, y(x))$.

Definition 10 *The numerical scheme (57) is said to be* **consistent** *with the differential equation (41) if the truncation error defined by (61) is such that for any* $\epsilon > 0$*, there exists an* $h(\epsilon)$ *for which*

$$|T_n| < \epsilon \quad \textit{for } 0 < h < h(\epsilon)$$

and any $(k+1)$ *points* $(x_n, y(x_n))$*,...,*$(x_{n+k}, y(x_{n+k}))$ *on any solution curve in* R *of the initial value problem (41), (42).*

Now let us suppose that the solution to the differential equation is sufficiently smooth, and let us expand $y(x_{n+j})$ and $y'(x_{n+j})$ into a Taylor series about the point x_n and substitute these expansions into the numerator in (61) to obtain

$$T_n = \frac{1}{h\sigma(1)}[C_0 y(x_n) + C_1 h y'(x_n) + C_2 h^2 y''(x_n) + \dots] \tag{62}$$

where

$$\begin{aligned} C_0 &= \sum_{j=0}^{k} \alpha_j \\ C_1 &= \sum_{j=1}^{k} j\alpha_j - \sum_{j=0}^{k} \beta_j \\ C_2 &= \sum_{j=1}^{k} \frac{j^2}{2!}\alpha_j - \sum_{j=1}^{k} j\beta_j \\ &\text{etc.} \\ C_q &= \sum_{j=1}^{k} \frac{j^q}{q!}\alpha_j - \sum_{j=1}^{k} \frac{j^{q-1}}{(q-1)!}\beta_j. \end{aligned}$$

For consistency we need that $T_n \to 0$ as $h \to 0$ and this requires that $C_0 = 0$ and $C_1 = 0$; in terms of the characteristic polynomials this consistency requirement can be restated in compact form as

$$\rho(1) = 0 \quad \text{and} \quad \rho'(1) = \sigma(1) \neq 0.$$

Let us observe that, according to this condition, if a linear multi-step method is consistent then it has a *simple* root on the unit circle at $z = 1$; thus the root condition is not violated by this zero.

Definition 11 *The numerical method (57) is said to have* **order of accuracy** p*, if* p *is the largest positive integer such that, for any sufficiently smooth solution curve in* R *of the initial value problem (41), (42), there exist constants* K *and* h_0 *such that*

$$|T_n| \leq Kh^p \quad \textit{for } 0 < h \leq h_0$$

for any $(k+1)$ *points* $(x_n, y(x_n))$,...,$(x_{n+k}, y(x_{n+k}))$ *on the solution curve.*

Thus we deduce from (62) that the method is of order of accuracy p if and only if

$$C_0 = C_1 = ... = C_p = 0 \quad \text{and} \quad C_{p+1} \neq 0.$$

In this case,

$$T_n = \frac{C_{p+1}}{\sigma(1)} h^p y^{(p+1)}(x_n) + O(h^{p+1}).$$

Exercise 12 (Oxford Finals, 1992) *Determine all values of the real parameter* b *for which the linear multi-step method*

$$y_{n+3} + (2b-3)(y_{n+2} - y_{n+1}) - y_n = hb(f_{n+2} + f_{n+1})$$

is zero-stable. Show that there exists a value of b *for which the order of the method is* 4*. Is the method convergent for this value of* b*? Show further that if the method is zero-stable than its order cannot exceed* 2*.*

SOLUTION: According to the root condition, this linear multi-step method is zero-stable if and only if all roots of its first characteristic polynomial

$$\rho(z) = z^3 + (2b-3)(z^2 - z) - 1$$

belong to the closed unit disc, and those on the unit circle are simple.

Clearly, $\rho(1) = 0$; upon dividing $\rho(z)$ by $z - 1$ we see that $\rho(z)$ can be written in the following factorised form:

$$\rho(z) = (z-1)\left(z^2 - 2(1-b)z + 1\right) \equiv (z-1)\rho_1(z).$$

Thus the method is zero stable if and only if all roots of the polynomial $\rho_1(z)$ belong to the closed unit disc, and those on the unit circle are simple and differ from 1. Suppose that the method is zero-stable. Then, it follows that $b \neq 0$ and $b \neq 2$, since these values of b correspond to double roots of $\rho_1(z)$ on the unit circle, respectively, $z = 1$ and $z = -1$. Since the product of the two roots of $\rho_1(z)$ is equal to 1 and neither of them is equal to ± 1, it follows that they must be strictly complex; hence the discriminant of the quadratic polynomial $\rho_1(z)$ must be negative. Namely,

$$4(1-b)^2 - 4 < 0.$$

In other words, $b \in (0, 2)$.

Conversely, suppose that $b \in (0, 2)$. Then the roots of $\rho(z)$ are

$$z_1 = 1, \qquad z_{2/3} = 1 - b + \imath\sqrt{1 - (b-1)^2}.$$

Since $|z_{2/3}| = 1$, $z_{2/3} \neq 1$ and $z_2 \neq z_3$, all roots of $\rho(z)$ lie on the unit circle and they are simple. Hence the method is zero-stable.

To summarise, the method is zero-stable if and only if $b \in (0, 2)$.

In order to analyse the order of accuracy of the method, we note that, upon Taylor series expansion, its truncation error can be written in the form

$$T_n = \left(1 - \frac{b}{6}\right) h^2 y'''(x_n) + \frac{1}{4}(6-b)h^3 y^{IV}(x_n) + \frac{1}{120}(150 - 23b)h^4 y^{V}(x_n) + O(h^5).$$

If $b = 6$, then $T_n = O(h^4)$ and so the method is of order 4. As $b = 6$ does not belong to the interval $(0, 2)$, we deduce that the method is *not* zero-stable for $b = 6$.

Since zero-stability requires $b \in (0, 2)$, in which case $1 - \frac{b}{6} \neq 0$, it follows that if the method is zero stable then $T_n = O(h^2)$. $\diamond$

We conclude this section with a convergence result for linear multi-step methods; this fundamental theorem was proved by the Swedish mathematician G. Dahlquist.

Theorem 21 (**Dahlquist**) *For a linear multi-step method that is consistent with the ordinary differential equation (41) where f is assumed to satisfy a Lipschitz condition, and starting with consistent initial data, zero stability is necessary and sufficient for convergence. Moreover if the solution $y(x)$ has continuous derivative of order $(p+1)$ and truncation error $O(h^p)$, then the global error $e_n = y(x_n) - y_n$ is also $O(h^p)$.*

For a proof of this result, see K.W. Morton, *Numerical Solution of Ordinary Differential Equations*, Oxford University Computing Laboratory, 1987, or P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962. By virtue of Dahlquist's theorem, if a linear multi-step method is not zero stable its global error cannot be made arbitrarily small by taking the mesh size h sufficiently small for any sufficiently accurate initial data. In fact, if the root condition is violated then there exists a solution to the linear multi-step method which will grow by an arbitrarily large factor in a fixed interval of x, however accurate the starting conditions are. This result highlights the importance of the concept of zero-stability and indicates its relevance in practical computations.

Further exercises

1. The values y_k of a cubic polynomial at the points $x_k = k$, $k = 0, ..., 5$, are given in the table below

$k = x_k$	0	1	2	3	4	5	6	7
y_k	1	2	4	8	15	26		

   Find the two missing values of y_k by constructing the Lagrange interpolation polynomial $p_3(x)$.

2. a) Show that the sequence $(a_n)_{n\geq 1}$, where

$$a_n = \left(\begin{array}{c} 2n \\ n \end{array} \right) \frac{1}{2^{2n+1}(n+1)},$$

   is monotonic decreasing. Show further, using Stirling's formula, that $\lim_{n\to\infty} a_n = 0$.

   b) Suppose that the function $f : x \mapsto \sqrt{x}$ has been tabulated on the interval $[1, 2]$ at equally spaced points $x_k = 1 + kh$, $k = 0, ..., n$, where $h = 1/n$. Show that

$$\max_{x\in[1,2]} |f(x) - p_n(x)| \leq a_n, \qquad \text{for } n \geq 1,$$

   where $p_n(x)$ is the Lagrange interpolation polynomial of degree n for f with interpolation points x_k, $k = 0, ..., n$. Deduce that $\lim_{n\to\infty} |f(x) - p_n(x)| = 0$ for each x in $[1, 2]$.

   Find $n_0 \geq 1$ such that

$$\max_{x\in[1,2]} |f(x) - p_n(x)| < 2 \times 10^{-2}$$

   for all $n \geq n_0$.

3. A switching path between parallel railway tracks can be described as a cubic polynomial joining $(0, 0)$ and $(4, 2)$ and tangent to the lines $y = 0$ and $y = 2$. Apply Hermite interpolation to construct this polynomial.

4. a) Show that

$$I = \int_0^1 \frac{1}{1+x^2} \mathrm{d}x = \frac{1}{4}\pi \ (\approx 0.78539816).$$

   b) Calculate the integral approximately, by subdividing the integral $[0, 1]$ into 10 sub-intervals, using a) the composite trapezium rule, b) the composite Simpson rule, and estimate the corresponding approximation errors.

5. Calculate the integral

$$\int_1^{1.5} \frac{1}{1+x^3} \mathrm{d}x,$$

   with error less then 10^{-4}, using the composite trapezium rule.

6. Suppose that
$$F(y) = \int_0^1 \tan^{-1} \mathrm{e}^{x^2+y^2+y}\,\mathrm{d}x.$$
Find $\min_{y>0} F(y)$ with accuracy $\epsilon = 10^{-6}$.

7. Suppose that
$$F(y) = \int_0^y \frac{\sin x}{x(2\pi - x)}\,\mathrm{d}x.$$
Find $\max_{y\in[0,3\pi]} F(y)$ with accuracy $\epsilon = 10^{-4}$.

8. Find the smallest positive solution of the equation
$$\int_1^x \frac{t}{\sin t}\,\mathrm{d}t = 2$$
with accuracy $\epsilon = 10^{-3}$.

9. Evaluate the integral
$$\int_{\pi/4}^{3\pi/4} \frac{\sin x}{x}\,\mathrm{d}x$$
with accuracy $\epsilon = 10^{-5}$.

10. Evaluate the integral
$$\int_1^5 \sqrt{1+\sqrt{x-1}}\,\mathrm{d}x$$
with accuracy $\epsilon = \frac{1}{2}10^{-2}$.

11. (Oxford Finals, 1997) Suppose that f is a real-valued function, defined and continuous on the closed real interval $[a, b]$. Write down the composite trapezium rule approximation $\mathcal{I}_m(f)$, with $m + 1$ quadrature points in the interval $[a, b]$, to the definite integral
$$\int_a^b f(x)\,\mathrm{d}x,$$
assuming that the spacing between consecutive quadrature points is $(b-a)/m$.

Suppose, further, that f has continuous second derivative f'' on the interval $[a, b]$. Show that
$$\left|\int_a^b f(x)\,\mathrm{d}x - \mathcal{I}_m(f)\right| \le \frac{1}{12m^2}(b-a)^3 \max_{x\in[a,b]} |f''(x)|.$$

Given that $\epsilon = 0.05$ and $f(x) = \tan^{-1} x$, for x in $[0, 1]$, find m, as small as possible, such that
$$\left|\int_0^1 f(x)\,\mathrm{d}x - \mathcal{I}_m(f)\right| \le \epsilon,$$
and compute $\mathcal{I}_m(f)$ for this value of m.

12. a) Construct the minimax polynomial of degree 1 for the function $f(x) = \log_2(1+x)$ on the interval $[0,1]$.

    b) Construct the minimax polynomial $p_1 \in \mathcal{P}_1$ for the function $g(x) = \sin(\pi x)$ on the interval $[-1,1]$. Prove that p_1 is, simultaneously, the minimax polynomial from $\mathcal{P}_2$ for the function g.

13. Suppose that $f \in C[-1,1]$ and let $f(-x) = -f(x)$ (respectively $f(-x) = f(x)$) for all x in $[-1,1]$. Further, let p_n be the minimax polynomial from $\mathcal{P}_n$ for the function f on the interval $[-1,1]$. Show that $p_n(-x) = -p_n(x)$ (respectively $p_n(-x) = p_n(x)$) for all $x \in [-1,1]$.

    Let $g(x) = \sin x$ for $x \in [-1,1]$. Find the minimax polynomial $p_2 \in \mathcal{P}_2$ for g on the interval $[-1,1]$.

    Let $h(x) = \cos x^2$ for $x \in [-1,1]$. Find the minimax polynomial $p_3 \in \mathcal{P}_3$ for h on the interval $[-1,1]$.

14. Among all polynomials of the form

$$p_n(x) = Ax^n + \sum_{k=0}^{n-1} a_k x^k,$$

    where $A \neq 0$ is a fixed real number, find the polynomial of best approximation to the function $f(x) \equiv 0$ on the interval $[-1,1]$.

15. Find the minimax polynomial $p_n \in \mathcal{P}_n$ for the function

$$f(x) = P_{n+1}(x) \equiv \sum_{k=0}^{n+1} a_k x^k, \qquad a_{n+1} \neq 0,$$

    on the interval $[-1,1]$.

16. Suppose that f is a continuous real-valued function on the closed interval $[a,b]$ of the real line. Suppose, further, that

$$\frac{f(x)-f(a)}{x-a} \leq \frac{f(b)-f(a)}{b-a}$$

    for all x in $(a,b]$. Show that there exists a set of critical points of the form $\{a,d,b\}$, with $d \in (a,b)$, for the minimax approximation of f by polynomials from $\mathcal{P}_1$.

17. Construct an example of a continuous function f defined on a closed interval $[a,b]$ such that the set of critical points for the minimax approximation of f by polynomials from $\mathcal{P}_1$ does not contain the points a and b.

18. Let $f(x) = |x|$ for $x \in [-1,2]$. Construct the minimax polynomial p_1 from $\mathcal{P}_1$ for f on the interval $[-1,2]$.

19. a) Prove Lemma 3 in the Lecture Notes.

    b) Find the minimax polynomial $p_5 \in \mathcal{P}_5$ for the function $f(x) = x^6$ on the interval $[-1, 1]$.

20. Construct the polynomial $p_3 \in \mathcal{P}_3$ of best approximation in the 2-norm for the function $x \mapsto \sin x$ on the interval $[-\pi, \pi]$, assuming that the weight function is $w(x) \equiv 1$.

21. Construct a system of orthogonal polynomials $\{\psi_0(x), \psi_1(x), \psi_2(x)\}$ on the interval $[-1, 1]$ with respect to the weight function $w(x) = x^2$. Determine the polynomial $p_2 \in \mathcal{P}_2$ of best approximation for the function $f(x) = x^4$ in the 2-norm $\|\cdot\|_2$ defined by

$$\|g\|_2 = \left(\int_{-1}^{1} w(x)|g(x)|^2 \mathrm{d}x \right)^{1/2}.$$

22. Suppose that $f(x) = \mathrm{e}^x$ for $x \in [0, 1]$. Find the polynomial $p_2 \in \mathcal{P}_2$ of best approximation for f in the 2-norm on the interval $[0, 1]$ with weight function $w(x) \equiv 1$.

23. (Oxford Finals, 1997) Given that m is a non-negative integer, let $\mathcal{P}_m$ denote the set of all polynomials of degree less than or equal to m. Construct the following polynomial approximations to the function $f : x \mapsto x^4$ on the interval $[-1, 1]$:

    a) the Hermite polynomial p in $\mathcal{P}_3$, which interpolates f at the points -1 and 1;

    b) the minimax polynomial q in $\mathcal{P}_3$ for f on the interval $[-1, 1]$;

    c) the best least-squares approximation r in $\mathcal{P}_3$ for f on the interval $[-1, 1]$, with respect to the weight function $w(x) \equiv 1$.

    Which of the three polynomials p, q, r is the least accurate approximation to f in the norm $\|\cdot\|_\infty$, defined by $\|f\|_\infty = \max_{x \in [-1,1]} |f(x)|$?

24. a) Find a non-negative integer n, as large as possible, and real numbers x_0 and A_0 such that

$$\int_{-1}^{1} x^2 f(x) \mathrm{d}x = A_0 f(x_0)$$

    whenever $f \in \mathcal{P}_{2n+1}$.

    b) Find a non-negative integer n, as large as possible, and real numbers x_1, x_2, A_1 and A_2 such that

$$\int_{-1}^{1} x^2 f(x) \mathrm{d}x = A_1 f(x_1) + A_2 f(x_2)$$

    whenever $f \in \mathcal{P}_{2n+1}$.

c) With x_i and A_i, $i = 0, 1, 2$, as in a) and b), consider the Gauss quadrature rules

$$\begin{aligned} \int_{-1}^{1} x^2 f(x)\mathrm{d}x &\approx A_0 f(x_0), \\ \int_{-1}^{1} x^2 f(x)\mathrm{d}x &\approx A_1 f(x_1) + A_2 f(x_2). \end{aligned}$$

Apply Theorem 10 from the Lecture Notes to estimate the size of the error for each of these quadrature rules.

25. On the interval $[0, 1]$ integrals of the form

$$\int_0^1 f(x)\sqrt{\frac{x}{a}}\,\mathrm{d}x, \qquad a > 0,$$

are approximated by the expression

$$Af(0) + Bf(1) + f(\xi),$$

where $\xi \in (0, 1)$. For what values of a is it possible to ensure that this approximation is exact for all polynomials from $\mathcal{P}_2$? Determine A, B and ξ in terms of a.

26. a) Consider the function $f(x) = \sin \pi x$, $x \in [0, 1]$, and suppose that $z_i = ih$, $i = 0, ..., m$, are equally spaced knots in $[0, 1]$, with $h = 1/m$, $m \geq 1$. Construct the linear spline $s_L(x)$ that interpolates the function f at these knots. Find m, as small as possible, such that

$$\|f - s_L\|_\infty \leq 0.1.$$

b) Now suppose that $f(x) = \sin \pi x$, $x \in [0, 1]$ is interpolated by a Hermite cubic spline $s(x)$ with the same knots as in part a). How large should m be to ensure that

$$\|f - s\|_\infty \leq 10^{-5}?$$

27. Write down Euler's method for the numerical solution of the initial value problem $y' + 5y = x\mathrm{e}^{-5x}$, $y(0) = 0$, on the interval $[0, 1]$ with step size $h = 1/N$, $N \geq 1$. Denoting by y_N the Euler approxiation to $y(1)$ at $x = 1$, show that $\lim_{N\to\infty} y_N = y(1)$. Find an integer N_0 such that

$$|y(1) - y_N| \leq 10^{-5}, \qquad \text{for all } N \geq N_0.$$

28. Consider the following one-step method for the numerical solution of the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$:

$$y_{n+1} = y_n + \frac{1}{2}h(k_1 + k_2),$$

where

$$k_1 = f(x_n, y_n), \qquad k_2 = f(x_n + h, y_n + hk_1).$$

Show that the method is consistent and has truncation error

$$T_n = \frac{1}{6}h^2 \left[f_y(f_x + f_y f) - \frac{1}{2}(f_{xx} + 2f_{xy}f + f_{yy}f^2) \right] + O(h^3).$$

29. Consider the one-step method

$$y_{n+1} = y_n + h(a\, k_1 + b\, k_2),$$

where

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + \alpha h, y_n + \beta h k_1), \end{aligned}$$

and where a, b, α, β are real parameters. Show that there is a choice of these parameters such that the order of the method is 2. Is there a choice of the parameters for which the order exceeds 2?

30. Consider the one-step method

$$y_{n+1} = y_n + \alpha h f(x_n, y_n) + \beta h f(x_n + \gamma h, y_n + \gamma h f(x_n, y_n)),$$

where α, β and γ are real parameters. Show that the method is consistent if and only if $\alpha + \beta = 1$. Show also that the order of the method cannot exceed 2. Suppose that a second-order method of the above form is applied to the initial value problem $y' = -\lambda y$, $y(0) = 1$, where λ is a positive real number. Show that the sequence $(y_n)_{n \geq 0}$ is bounded if and only if $h \leq \frac{2}{\lambda}$. Show further that, for such λ,

$$|y(x_n) - y_n| \leq \frac{1}{6}\lambda^3 h^2 x_n, \quad n \geq 0.$$

31. Consider the linear two-step method

$$y_{n+2} - y_n = \frac{h}{3}(f_{n+2} + 4f_{n+1} + f_n).$$

Show that the method is zero stable; show further that it is third-order accurate, namely, $T_n = O(h^3)$.

32. Show that the linear three-step method

$$11y_{n+3} + 27y_{n+2} - 27y_{n+1} - 11y_n = 3h[f_{n+3} + 9f_{n+2} + 9f_{n+1} + f_n]$$

is sixth order accurate. Find the roots of the first characteristic polynomial and deduce that the method is not zero-stable.

33. (Oxford Finals, 1997) Write down the general form of a linear multi-step method for the numerical solution of the initial value problem

$$y' = f(x, y), \qquad y(x_0) = y_0,$$

on the closed real interval $[x_0, x_N]$, where f is a continuous function of its arguments and y_0 is a given real number. What is meant by saying that the method is *zero-stable*? Define the *truncation error* of the method. What does it mean to say that the method has *order of accuracy* p?

Given that α is a positive real number, consider the linear two-step method

$$y_{n+2} - \alpha y_n = \frac{h}{3}\left[f(x_{n+2}, y_{n+2}) + 4f(x_{n+1}, y_{n+1}) + f(x_n, y_n)\right],$$

on the mesh $\{x_n : x_n = x_0 + nh, n = 1, ..., N\}$ of spacing h, $h > 0$. Determine the set of all α such that the method is zero-stable. Find α such that the order of accuracy is as high as possible; is the method convergent for this value of α?