Chapter 3

The Minimax Property and Its Applications

3.1 Approximation — theory and structure

One area above all in which the Chebyshev polynomials have a pivotal role is the minimax approximation of functions by polynomials. It is therefore appropriate at the beginning of this discussion to trace the structure of the subject of approximation and to present some essential theoretical results, concentrating primarily on uniform (or $\mathcal{L}_\infty$) approximation and introducing the minimax property of the Chebyshev polynomials.

It is very useful to be able to replace any given function by a simpler function, such as a polynomial, chosen to have values not identical with but very close to those of the given function, since such an 'approximation' may not only be more compact to represent and store but also more efficient to evaluate or otherwise manipulate. The structure of an 'approximation problem' involves three central components: (i) a *function class* (containing the function to be approximated), (ii) a *form* (for the approximating function) and (iii) a *norm* (of the approximation error), in terms of which the problem may be formally posed. The expert's job is to make appropriate selections of these components, then to pose the approximation problem, and finally to solve it.

By a *function class*, we mean a restricted family $\mathcal{F}$ of functions f to which any function $f(x)$ that we may want to fit is assumed to belong. Unless otherwise stated, we shall be concerned with real functions of a real variable, but the family will generally be narrower than this. For example we may consider amongst others the following alternative families $\mathcal{F}$ of functions defined on the real interval $[a, b]$:

1. $\mathcal{C}[a, b]$: continuous functions on $[a, b]$;

2. $\mathcal{L}_\infty[a, b]$: bounded functions on $[a, b]$;

3. $\mathcal{L}_2[a, b]$: square-integrable functions on $[a, b]$;

4. $\mathcal{L}_p[a, b]$: $\mathcal{L}_p$-integrable functions on $[a, b]$, namely functions $f(x)$ for which is defined

$$\int_a^b w(x)\,|f(x)|^p\,\mathrm{d}x, \tag{3.1}$$

where $w(x)$ is a given non-negative weight function and $1 \le p < \infty$. Note that $\mathcal{L}_2[a, b]$ is a special case ($p = 2$) of $\mathcal{L}_p[a, b]$.

© 2003 by CRC Press LLC

The reason for defining such a family of functions, when in practice we may only in fact be interested in one specific function, is that this helps to isolate those properties of the function that are relevant to the theory — moreover, there is a close link between the function class we work in and the norms we can use. In particular, in placing functions in one of the four families listed above, it is implicitly assumed that we neither care how the functions behave nor wish to approximate them outside the given interval $[a, b]$.

By *form of approximation* we mean the specific functional form which is to be adopted, which will always include adjustable coefficients or other parameters. This defines a family $\mathcal{A}$ of possible approximations $f^*(x)$ to the given function $f(x)$. For example, we might draw our approximation from one of the following families:

1. Polynomials of degree n, with

$$\mathcal{A} = \Pi_n = \{f^*(x) = p_n(x) = c_0 + c_1 x + \cdots + c_n x^n\} \quad \text{(parameters } \{c_j\})$$

2. Rational functions of type (p, q), with

$$\mathcal{A} = \left\{ f^*(x) = r_{p,q}(x) = \frac{a_0 + a_1 + \cdots + a_p x^p}{1 + b_1 + \cdots + b_q x^q} \right\} \quad \text{(parameters } \{a_j\}, \{b_j\})$$

For theoretical purposes it is usually desirable to choose the function class $\mathcal{F}$ to be a *vector space* (or linear space). A vector space $\mathcal{V}$ comprises elements u, v, w, ... with the properties (which vectors in the conventional sense are easily shown to possess):

1. (*closure under addition*)
   $u + v \in \mathcal{V}$ for any $u, v \in \mathcal{V}$,

2. (*closure under multiplication by a scalar*)
   $\alpha u \in \mathcal{V}$ for any $u \in \mathcal{V}$ and for any scalar α.

When these elements are functions $f(x)$, with $f + g$ and αf defined as the functions whose values at any point x are $f(x) + g(x)$ and $\alpha f(x)$, we refer to $\mathcal{F}$ as a *function space*. This space $\mathcal{F}$ typically has infinite dimension, the 'vector' in question consisting of the values of $f(x)$ at each of the continuum of points x in $[a, b]$.

The family $\mathcal{A}$ of approximations is normally taken to be a subclass of $\mathcal{F}$:

$$\mathcal{A} \subset \mathcal{F}$$

— in practice, $\mathcal{A}$ is usually also a vector space, and indeed a function space. In contrast to $\mathcal{F}$, $\mathcal{A}$ is a *finite dimensional* function space, its dimension being the number of parameters in the form of approximation. Thus the space Π_n of

© 2003 by CRC Press LLC

polynomials $p_n(x)$ of degree n has dimension $n+1$ and is in fact isomorphic (i.e., structurally equivalent) to the space $\mathbb{R}^{n+1}$ of real vectors with $n+1$ components:

$$\{\mathbf{c} = (c_0, c_1, \ldots, c_n)\}.$$

(Note that the family of rational functions $r_{p,q}$ of type (p,q) is *not* a vector space, since the sum of two such functions is in general a rational function of type $(p+q, 2q)$, which is not a member of the same family.)

The *norm* of approximation $\|\cdot\|$ serves to compare the function $f(x)$ with the approximation $f^*(x)$, and gives a single scalar measure of the closeness of f^* to f, namely

$$\|f - f^*\|. \tag{3.2}$$

Definition 3.1 *A* norm $\|\cdot\|$ *is defined as any real scalar measure of elements of a vector space that satisfies the axioms:*

1. $\|u\| \geq 0$, *with equality if and only if* $u \equiv 0$*;*
2. $\|u+v\| \leq \|u\| + \|v\|$ *(*the 'triangle inequality'*);*
3. $\|\alpha u\| = |\alpha|\ \|u\|$ *for any scalar* α*.*

Such a definition encompasses all the key features of distance or, in the case of a function, size. Standard choices of norm for function spaces are the following:

1. $\mathcal{L}_\infty$ norm (or *uniform norm*, *minimax norm*, or *Chebyshev norm*):
$$\|f\| = \|f\|_\infty = \max_{a \leq x \leq b} |f(x)|; \tag{3.3}$$
2. $\mathcal{L}_2$ norm (or *least-squares norm*, or *Euclidean norm*):
$$\|f\| = \|f\|_2 = \sqrt{\int_a^b w(x)\,|f(x)|^2\, \mathrm{d}x}, \tag{3.4}$$
where $w(x)$ is a non-negative weight function;
3. $\mathcal{L}_1$ norm (or *mean norm*, or *Manhattan norm*):
$$\|f\| = \|f\|_1 = \int_a^b w(x)\,|f(x)|\, \mathrm{d}x; \tag{3.5}$$
4. The above three norms can be collected into the more general $\mathcal{L}_p$ norm (or *Hölder norm*):
$$\|f\| = \|f\|_p = \left[\int_a^b w(x)\,|f(x)|^p\, \mathrm{d}x\right]^{\frac{1}{p}}, \quad (1 \leq p < \infty), \tag{3.6}$$

© 2003 by CRC Press LLC

where $w(x)$ is a non-negative weight function.

With suitable restrictions on f, which are normally satisfied in practice, this $\mathcal{L}_p$ norm corresponds to the $\mathcal{L}_\infty$, $\mathcal{L}_2$ and $\mathcal{L}_1$ norms in the cases $p \to \infty$, $p = 2$, $p = 1$, respectively.

5. The *weighted minimax norm*:

$$\|f\| = \max_{a \leq x \leq b} w(x) \, |f(x)| \tag{3.7}$$

(which does not fall into the pattern of Hölder norms) also turns out to be appropriate in some circumstances.

The $\mathcal{L}_p$ norm becomes stronger as p increases, as the following lemma indicates.

Lemma 3.1 *If $1 \leq p_1 < p_2 \leq \infty$, and if a, b and $\int_a^b w(x)\,\mathrm{d}x$ are finite, then $\mathcal{L}_{p_2}[a,b]$ is a subspace of $\mathcal{L}_{p_1}[a,b]$, and there is a finite constant $k_{p_1 p_2}$ such that*

$$\|f\|_{p_1} \leq k_{p_1 p_2} \|f\|_{p_2} \tag{3.8}$$

for every f in $\mathcal{L}_{p_2}[a,b]$.

This lemma will be deduced from Hölder's inequality in Chapter 5 (see Lemma 5.4 on page 117).

A vector space to which a norm has been attached is termed a *normed linear space*. Hence, once a norm is chosen, the vector spaces $\mathcal{F}$ and $\mathcal{A}$ of functions and approximations become normed linear spaces.

3.1.1 The approximation problem

We defined above a family of functions or function space, $\mathcal{F}$, a family of approximations or approximation (sub)space, $\mathcal{A}$, and a measure $\|f - f^*\|$ of how close a given function $f(x)$ in $\mathcal{F}$ is to a derived approximation $f^*(x)$ in $\mathcal{A}$. How then do we more precisely judge the quality of $f^*(x)$, as an approximation to $f(x)$ in terms of this measure? In practice there are three types of approximation that are commonly aimed for:

Definition 3.2 *Let $\mathcal{F}$ be a normed linear space, let $f(x)$ in $\mathcal{F}$ be given, and let $\mathcal{A}$ be a given subspace of $\mathcal{F}$.*

1. *An approximation $f^*(x)$ in $\mathcal{A}$ is said to be* good *(or* acceptable*) if*

$$\|f - f^*\| \leq \epsilon \tag{3.9}$$

where ϵ is a prescribed level of absolute accuracy.

© 2003 by CRC Press LLC

2. *An approximation $f_B^*(x)$ in $\mathcal{A}$ is a* best *approximation if, for any other approximation $f^*(x)$ in $\mathcal{A}$,*

$$\|f - f_B^*\| \leq \|f - f^*\| . \tag{3.10}$$

   Note that there will sometimes be more than one best approximation to the same function.

3. *An approximation $f_N^*(x)$ in $\mathcal{A}$ is said to be* near-best within a relative distance *ρ if*

$$\|f - f_N^*\| \leq (1 + \rho) \|f - f_B^*\| , \tag{3.11}$$

   where ρ is a specified positive scalar and $f_B^*(x)$ is a best approximation.

In the case of the $\mathcal{L}_\infty$ norm, we often use the terminology minimax *and* near-minimax *in place of* best *and* near-best*.*

The 'approximation problem' is to determine an approximation of one of these types (good, best or near-best). In fact, it is commonly required that both 'good' and 'best', or both 'good' and 'near-best', should be achieved — after all, it cannot be very useful to obtain a best approximation if it is also a very poor approximation.

In defining 'good' in Definition 3.2 above, an absolute error criterion is adopted. It is, however, also possible to adopt a *relative error* criterion, namely

$$\left\| 1 - \frac{f^*}{f} \right\| \leq \epsilon. \tag{3.12}$$

This can be viewed as a problem of weighted approximation in which we require

$$\|w\ (f - f^*)\| \leq \epsilon, \tag{3.13}$$

where, in this case,

$$w(x) = 1/ |f(x)| .$$

In approximating by polynomials on $[a, b]$, it is always possible to obtain a good approximation by taking the degree high enough. This is the conclusion of the following well-known results.

Theorem 3.2 (Weierstrass's theorem) *For any given f in $\mathcal{C}[a, b]$ and for any given $\epsilon > 0$, there exists a polynomial p_n for some sufficiently large n such that $\|f - p_n\|_\infty < \epsilon$.*

Proof: A proof of this will be given later (see Corollary 5.8A on page 120). ●●

Corollary 3.2A *The same holds for $\|f - p_n\|_p$ for any $p \geq 1$.*

© 2003 by CRC Press LLC

Proof: This corollary follows directly by applying Lemma 3.1. ●●

But of course it is a good thing from the point of view of efficiency if we can keep the degree of polynomial as low as possible, which we can do by concentrating on best or near-best approximations.

3.2 Best and minimax approximation

Given a norm $\|\cdot\|$ (such as $\|\cdot\|_\infty$, $\|\cdot\|_2$ or $\|\cdot\|_1$), a best approximation as defined by (3.10) is a solution of the problem:

$$\underset{f^* \in \mathcal{A}}{\text{minimise}} \|f - f^*\| . \tag{3.14}$$

In the case of polynomial approximation:

$$f^*(x) = p_n(x) = c_0 + c_1 x + \cdots + c_n x^n, \tag{3.15}$$

to which we now restrict our attention, we may rewrite (3.14) in terms of the parameters as:

$$\underset{c_0,\ldots,c_n}{\text{minimise}} \|f - p_n\| . \tag{3.16}$$

Can we always find such a p_n? Is there just one?

Theorem 3.3 *For any given p $(1 \leq p \leq \infty)$, there exists a unique best polynomial approximation p_n to any function $f \in \mathcal{L}_p[a, b]$ in the $\mathcal{L}_p$ norm, where $w(x)$ is taken to be unity in the case $p \to \infty$.*

We refrain from giving proofs, but refer the reader to Cheney (1966), or other standard texts, for details.

Note that best approximations also exist in $\mathcal{L}_p$ norms on finite point sets, for $1 \leq p \leq \infty$, and are then unique for $p \neq 1$ but not necessarily unique for $p = 1$. Such $\mathcal{L}_p$ norms are defined by:

$$\|f - f^*\|_p = \left[\sum_{i=1}^{m} w_i \left| f(x_i) - f^*(x_i) \right|^p \right]^{\frac{1}{p}}$$

where $\{w_i\}$ are positive scalar weights and $\{x_i\}$ is a discrete set of m fitting points where the value of $f(x_i)$ is known. These are important in data fitting problems; however, this topic is away from our central discussion, and we shall not pursue it here.

It is possible to define forms of approximation other than polynomials, for which existence or uniqueness of best approximation holds — see Cheney (1966) for examples. Since polynomials are the subject of this book, however, we shall again refrain from going into details.

© 2003 by CRC Press LLC

Note that Theorem 3.3 guarantees in particular the existence of a unique best approximation in the $\mathcal{L}_\infty$ or minimax norm. The best $\mathcal{L}_\infty$ or minimax approximation problem, combining (3.3) and (3.15), is (in concise notation)

$$\underset{c_0,\ldots,c_n}{\text{minimise}} \max_{a \le x \le b} |f(x) - p_n(x)| \,. \tag{3.17}$$

It is clear from (3.17) why the word 'minimax' is often given to this problem, and why the resulting best approximation is often referred to as a 'minimax approximation'.

Theorem 3.3 is not a constructive theorem and does not characterise (i.e. describe how to recognise) a minimax approximation. However, it is possible to do so rather explicitly, as the following powerful theorem asserts.

Theorem 3.4 (Alternation theorem for polynomials) *For any $f(x)$ in $\mathcal{C}[a, b]$ a unique minimax polynomial approximation $p_n(x)$ exists, and is uniquely characterised by the 'alternating property' (or 'equioscillation property') that there are $n + 2$ points (at least) in $[a, b]$ at which $f(x) - p_n(x)$ attains its maximum absolute value (namely $\|f - p_n\|_\infty$) with alternating signs.*

This theorem, often ascribed to Chebyshev but more properly attributed to Borel (1905), asserts that, for p_n to be the best approximation, it is both *necessary and sufficient* that the alternating property should hold, that only one polynomial has this property, and that there is only one best approximation. The reader is referred to Cheney (1966), for example, for a complete proof. The 'sufficient' part of the proof is relatively straightforward and is set as Problem 6 below; the 'necessary' part of the proof is a little more tricky.

Example 3.1: As an example of the alternation theorem, suppose that the function $f(x) = x^2$ is approximated by the first-degree ($n = 1$) polynomial

$$f^*(x) = p_1(x) = x - 0.125 \tag{3.18}$$

on $[0, 1]$. Then the error $f(x) - p_n(x)$, namely

$$x^2 - x + 0.125,$$

has a maximum magnitude of 0.125 which it attains at $x = 0$, 0.5 and 1. At these points it takes the respective values +0.125, −0.125 and +0.125, which have alternating signs. (See Figure 3.1.) Hence $p_1(x)$, given by (3.18), is the unique minimax approximation.

Define $\mathcal{C}^0_{2\pi}$ to be the space of functions which are continuous and 2π-periodic (so that $f(2\pi+\theta) = f(\theta)$). There is a theorem similar to Theorem 3.4

© 2003 by CRC Press LLC

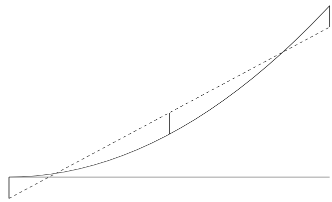

Figure 3.1: Minimax linear approximation to x^2 on range $[0, 1]$

which holds for approximation of a continuous function by a *trigonometric polynomial*, such as

$$q_n(\theta) = a_0 + \sum_{k=1}^{n} (a_k \cos k\theta + b_k \sin k\theta) \tag{3.19}$$

on the range $[-\pi, \pi]$ of θ.

Theorem 3.5 (Alternation theorem for trigonometric polynomials) *For any $f(\theta)$ in $\mathcal{C}^0_{2\pi}$, the minimax approximation $q_n(\theta)$ of form (3.19) exists and is uniquely characterised by an alternating property at $2n + 2$ points of $[-\pi, \pi]$. If $b_1, \ldots, b_n$ (or $a_0, \ldots, a_n$) are set to zero, so that $q_n(\theta)$ is a sum of cosine (or sine) functions alone, and if $f(\theta)$ is an even (or odd) function, then the minimax approximation $q_n(\theta)$ is characterised by an alternating property at $n + 2$ (or respectively $n + 1$) points of $[0, \pi]$.*

Finally, we should mention recent work by Peherstorfer (1997, and elsewhere) on minimax polynomial approximation over collections of non-overlapping intervals.

3.3 The minimax property of the Chebyshev polynomials

We already know, from our discussions of Section 2.2, that the Chebyshev polynomial $T_n(x)$ has $n + 1$ extrema, namely

$$x = y_k = \cos \frac{k\pi}{n} \quad (k = 0, 1, \ldots, n). \tag{3.20}$$

Since $T_n(x) = \cos n\theta$ when $x = \cos\theta$ (by definition), and since $\cos n\theta$ attains its maximum magnitude of unity with alternating signs at its extrema, the following property holds.

Lemma 3.6 (Alternating property of $T_n(x)$) *On $[-1, 1]$, $T_n(x)$ attains its maximum magnitude of 1 with alternating signs at precisely $(n + 1)$ points, namely the points (3.20).*

© 2003 by CRC Press LLC

Clearly this property has the flavour of the alternation theorem for minimax polynomial approximation, and indeed we can invoke this theorem as follows. Consider the function

$$f(x) = x^n,$$

and consider its minimax polynomial approximation of degree $n-1$ on $[-1, 1]$, $p_{n-1}(x)$, say. Then, by Theorem 3.4, $f(x) - p_{n-1}(x) = x^n - p_{n-1}(x)$ must uniquely have the alternating property on $n+1$ points. But $T_n(x)$ has a leading coefficient (of x^n) equal to 2^{n-1} and hence $2^{1-n}T_n(x)$ is of the same form $x^n - p_{n-1}(x)$ with the same alternating property. It follows that

$$x^n - p_{n-1}(x) = 2^{1-n}T_n(x). \tag{3.21}$$

We say that $2^{1-n}T_n(x)$ is a *monic polynomial*, namely a polynomial with unit leading coefficient. The following two corollaries of the alternation theorem now follow.

Corollary 3.4A (of Theorem 3.4) *The minimax polynomial approximation of degree $n-1$ to the function $f(x) = x^n$ on $[-1, 1]$ is*

$$p_{n-1}(x) = x^n - 2^{1-n}T_n(x). \tag{3.22}$$

Corollary 3.4B (The minimax property of T_n) $2^{1-n}T_n(x)$ *is the minimax approximation on $[-1, 1]$ to the zero function by a monic polynomial of degree n.*

EXAMPLE 3.2: As a specific example of Corollary 3.4B, the minimax monic polynomial approximation of degree $n = 4$ to zero on $[-1, 1]$ is

$$2^{-3}T_4(x) = 2^{-3}(8x^4 - 8x^2 + 1) = x^4 - x^2 + 0.125.$$

This polynomial has the alternating property, taking extreme values $+0.125$, -0.125, $+0.125$, -0.125, $+0.125$, respectively, at the 5 points $y_k = \cos k\pi/4$ $(k = 0, 1, \ldots, 4)$, namely

$$y_k = 1,\ \frac{1}{\sqrt{2}},\ 0,\ -\frac{1}{\sqrt{2}},\ -1. \tag{3.23}$$

Moreover, by Corollary 3.4A, the minimax cubic polynomial approximation to the function $f(x) = x^4$ on $[-1, 1]$ is, from (3.22),

$$p_3(x) = x^4 - (x^4 - x^2 + 0.125) = x^2 - 0.125, \tag{3.24}$$

the error $f(x) - p_3(x)$ having the alternating property at the points (3.23). Thus the minimax cubic polynomial approximation in fact reduces to a quadratic polynomial in this case.

It is noteworthy that $x^2 - 0.125$ is also the minimax quadratic polynomial $(n = 2)$ approximation to x^4 on $[-1, 1]$. The error still has 5 extrema, and so in this case the

© 2003 by CRC Press LLC

alternation theorem holds with $n+3$ alternation points. It is thus certainly possible for the number of alternation points to exceed $n+2$.

If the interval of approximation is changed to $[0,1]$, then a shifted Chebyshev polynomial is required. Thus the minimax monic polynomial approximation of degree n to zero on $[0,1]$ is

$$2^{1-2n}T_n^*(x). \tag{3.25}$$

For example, for $n=2$, the minimax monic quadratic is

$$2^{-3}T_2^*(x) = 2^{-3}(8x^2-8x+1) = x^2-x+0.125.$$

This is precisely the example (3.18) that was first used to illustrate Theorem 3.4 above.

3.3.1 Weighted Chebyshev polynomials of second, third and fourth kinds

We saw above that the minimax property of $T_n(x)$ depended on the alternating property of $\cos n\theta$. However, an alternating property holds at $n+1$ points θ in $[0,\pi]$ for each of the trigonometric polynomials

$$\begin{array}{lll}
\sin(n+1)\theta\,, & \text{at } \theta = \dfrac{(k+\frac{1}{2})\pi}{n+1} & (k=0,\ldots,n), \\
\cos(n+\frac{1}{2})\theta, & \text{at } \quad \theta = \dfrac{k\pi}{n+\frac{1}{2}} & (k=0,\ldots,n), \\
\sin(n+\frac{1}{2})\theta, & \text{at } \theta = \dfrac{(k+\frac{1}{2})\pi}{n+\frac{1}{2}} & (k=0,\ldots,n).
\end{array}$$

The following properties may therefore readily be deduced from the definitions (1.4), (1.8) and (1.9) of $U_n(x)$, $V_n(x)$, $W_n(x)$.

Corollary 3.5A (of Theorem 3.5) *(Weighted minimax properties of U_n, V_n, W_n)*

The minimax approximations to zero on $[-1,1]$, by monic polynomials of degree n weighted respectively by $\sqrt{1-x^2}$, $\sqrt{1+x}$ and $\sqrt{1-x}$, are

$$2^{-n}U_n(x),\ 2^{-n}V_n(x)\ \text{and}\ 2^{-n}W_n(x).$$

The characteristic equioscillation may be seen in Figure 3.2.

© 2003 by CRC Press LLC

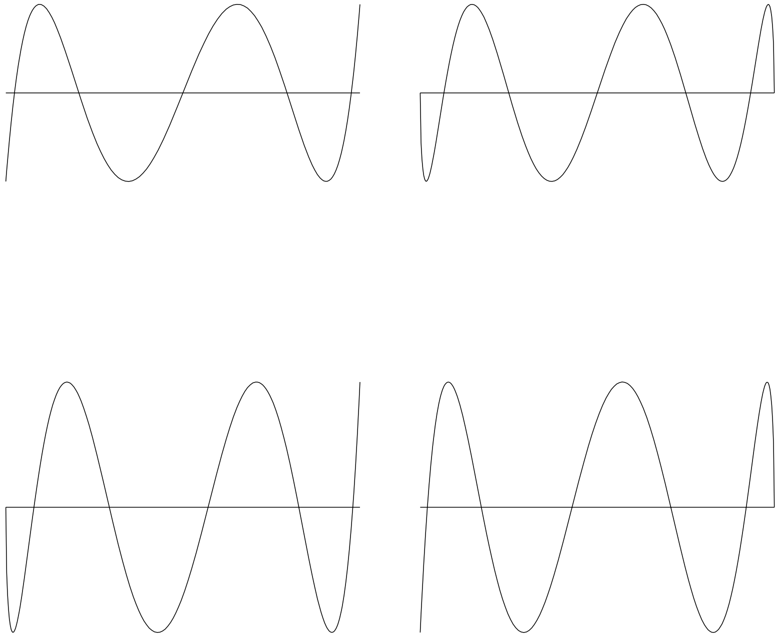

Figure 3.2: Equioscillation on $[-1, 1]$ of $T_5(x)$, $\sqrt{1-x^2}U_5(x)$, $\sqrt{1+x}V_5(x)$ and $\sqrt{1-x}W_5(x)$

© 2003 by CRC Press LLC

3.4 The Chebyshev semi-iterative method for linear equations

The minimax property of the Chebyshev polynomials T_n has been exploited to accelerate the convergence of iterative solutions of linear algebraic equations (Varga 1962, p.138), (Golub & van Loan 1983, p.511).

Let a set of linear equations be written in matrix form as

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{3.26}$$

Then a standard method of solution is to express the square matrix $\mathbf{A}$ in the form $\mathbf{A} = \mathbf{M} - \mathbf{N}$, where the matrix $\mathbf{M}$ is easily inverted (e.g., a diagonal or banded matrix), to select an initial vector $\mathbf{x}_0$, and to perform the iteration

$$\mathbf{M}\mathbf{x}_{k+1} = \mathbf{N}\mathbf{x}_k + \mathbf{b}. \tag{3.27}$$

This iteration will converge to the solution $\mathbf{x}$ of (3.26) if the *spectral radius* $\rho(\mathbf{G})$ of the matrix $\mathbf{G} = \mathbf{M}^{-1}\mathbf{N}$ (absolute value of its largest eigenvalue) is less than unity, converging at a geometric rate proportional to $\rho(\mathbf{G})^k$.

Now suppose that we replace each iterate $\mathbf{x}_k$ by a linear combination of successive iterates:

$$\mathbf{y}_k = \sum_{j=0}^{k} \nu_j(k)\mathbf{x}_j \tag{3.28}$$

where

$$\sum_{j=0}^{k} \nu_j(k) = 1, \tag{3.29}$$

and write

$$p_k(z) := \sum_{j=0}^{k} \nu_j(k) z^j,$$

so that $p_k(1) = 1$.

From (3.26) and (3.27), we have $\mathbf{M}(\mathbf{x}_{j+1} - \mathbf{x}) = \mathbf{N}(\mathbf{x}_j - \mathbf{x})$, so that

$$\mathbf{x}_j - \mathbf{x} = \mathbf{G}^j(\mathbf{x}_0 - \mathbf{x})$$

and, substituting in (3.28) and using (3.29),

$$\mathbf{y}_k - \mathbf{x} = \sum_{j=0}^{k} \nu_j(k)\mathbf{G}^j(\mathbf{x}_0 - \mathbf{x}) = p_k(\mathbf{G})(\mathbf{x}_0 - \mathbf{x}), \tag{3.30}$$

where $p_k(\mathbf{G})$ denotes the matrix $\sum_{j=0}^{k} \nu_j(k)\mathbf{G}^j$.

Assume that the matrix $\mathbf{G} = \mathbf{M}^{-1}\mathbf{N}$ has all of its eigenvalues $\{\lambda_i\}$ real and lying in the range $[\alpha, \beta]$, where $-1 < \alpha < \beta < +1$. Then $p_k(\mathbf{G})$ has eigenvalues $p_k(\lambda_i)$, and

$$\rho(p_k(\mathbf{G})) = \max_i |p_k(\lambda_i)| \leq \max_{\alpha \leq \lambda \leq \beta} |p_k(\lambda)|. \tag{3.31}$$

© 2003 by CRC Press LLC

Let F denote the linear mapping of the interval $[\alpha, \beta]$ onto the interval $[-1, 1]$:

$$F(z) = \frac{2z - \alpha - \beta}{\beta - \alpha} \tag{3.32}$$

and write

$$\mu = F(1) = \frac{2 - \alpha - \beta}{\beta - \alpha}. \tag{3.33}$$

Choose the coefficients $\nu_j(k)$ so that

$$p_k(z) = \frac{T_k(F(z))}{T_k(\mu)}. \tag{3.34}$$

Then $p_k(1) = 1$, as required, and

$$\max_{\alpha \le \lambda \le \beta} |p_k(\lambda)| = \frac{1}{|T_k(\mu)|} = \frac{1}{\cosh(k \operatorname{argcosh} \mu)} \sim 2\mathrm{e}^{-k \operatorname{argcosh} \mu}, \tag{3.35}$$

using (1.33a) here, rather than (1.1), since we know that $\mu > 1$. Convergence of $\mathbf{y}_k$ to $\mathbf{x}$ is therefore rapid, provided that μ is large.

It remains to show that $\mathbf{y}_k$ can be computed much more efficiently than by computing $\mathbf{x}_k$ and evaluating the entire summation (3.28) at every step. We can achieve this by making use of the recurrence (1.3a) in the forms

$$\begin{aligned} T_{k-1}(\mu) &= 2\mu\, T_k(\mu) - T_{k+1}(\mu) \\ T_{k+1}(\mathbf{\Gamma}) &= 2\mathbf{\Gamma}\, T_k(\mathbf{\Gamma}) - T_{k-1}(\mathbf{\Gamma}) \end{aligned} \tag{3.36}$$

where

$$\mathbf{\Gamma} = F(\mathbf{G}) = \frac{2}{\beta - \alpha}\mathbf{G} - \frac{\beta + \alpha}{\beta - \alpha}. \tag{3.37}$$

From (3.30) we have

$$\begin{aligned} \mathbf{y}_{k+1} - \mathbf{y}_{k-1} &= (\mathbf{y}_{k+1} - \mathbf{x}) - (\mathbf{y}_{k-1} - \mathbf{x}) \\ &= p_{k+1}(\mathbf{G})(\mathbf{x}_0 - \mathbf{x}) - p_{k-1}(\mathbf{G})(\mathbf{x}_0 - \mathbf{x}) \\ &= \left(\frac{T_{k+1}(\mathbf{\Gamma})}{T_{k+1}(\mu)} - \frac{T_{k-1}(\mathbf{\Gamma})}{T_{k-1}(\mu)} \right)(\mathbf{x}_0 - \mathbf{x}); \\ \mathbf{y}_k - \mathbf{y}_{k-1} &= \left(\frac{T_k(\mathbf{\Gamma})}{T_k(\mu)} - \frac{T_{k-1}(\mathbf{\Gamma})}{T_{k-1}(\mu)} \right)(\mathbf{x}_0 - \mathbf{x}). \end{aligned}$$

Define

$$\omega_{k+1} = 2\mu \frac{T_k(\mu)}{T_{k+1}(\mu)}. \tag{3.38}$$

Then, using (3.36), the expression

$$(\mathbf{y}_{k+1} - \mathbf{y}_{k-1}) - \omega_{k+1}(\mathbf{y}_k - \mathbf{y}_{k-1})$$

© 2003 by CRC Press LLC

simplifies to

$$\begin{aligned} 2(\mathbf{\Gamma}-\mu)\frac{T_k(\mathbf{\Gamma})}{T_{k+1}(\mu)}(\mathbf{x}_0-\mathbf{x}) &= \omega_{k+1}\frac{\mathbf{\Gamma}-\mu}{\mu}(\mathbf{y}_k-\mathbf{x}) \\ &= \omega_{k+1}\gamma(\mathbf{G}-1)(\mathbf{y}_k-\mathbf{x}) = \omega_{k+1}\gamma\mathbf{z}_k \end{aligned}$$

where

$$\gamma = 2/(2-\alpha-\beta) \tag{3.39}$$

and where $\mathbf{z}_k$ satisfies

$$\begin{aligned} \mathbf{M}\mathbf{z}_k &= \mathbf{M}(\mathbf{G}-1)(\mathbf{y}_k-\mathbf{x}) \\ &= (\mathbf{N}-\mathbf{M})(\mathbf{y}_k-\mathbf{x}) = \mathbf{A}(\mathbf{x}-\mathbf{y}_k) = \mathbf{b}-\mathbf{A}\mathbf{y}_k. \end{aligned} \tag{3.40}$$

The successive iterates $\mathbf{y}_k$ can thus be generated by means of the three-term recurrence

$$\mathbf{y}_{k+1} = \omega_{k+1}(\mathbf{y}_k - \mathbf{y}_{k-1} + \gamma\mathbf{z}_k) + \mathbf{y}_{k-1}, \quad k = 1, 2, \ldots, \tag{3.41}$$

starting from

$$\mathbf{y}_0 = \mathbf{x}_0, \quad \mathbf{y}_1 = \mathbf{y}_0 + \gamma\mathbf{z}_0, \tag{3.42}$$

where

$$\omega_{k+1} = 2\mu\frac{T_k(\mu)}{T_{k+1}(\mu)}, \quad \mu = \frac{2-\alpha-\beta}{\beta-\alpha}, \quad \gamma = \frac{2}{2-\alpha-\beta},$$

and $\mathbf{z}_k$ is at each step the solution of the linear system

$$\mathbf{M}\mathbf{z}_k = \mathbf{b} - \mathbf{A}\mathbf{y}_k. \tag{3.43}$$

Using (1.3a) again, we can generate the coefficients ω_k most easily by means of the recurrence

$$\omega_{k+1} = \frac{1}{1-\omega_k/4\mu^2} \tag{3.44}$$

with $\omega_1 = 2$; they converge to a limit $\omega_k \to 2\mu(\mu - \sqrt{\mu^2-1})$ as $k \to \infty$.

In summary, the algorithm is as follows:

Given the system of linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, with $\mathbf{A} = \mathbf{M}-\mathbf{N}$, where $\mathbf{M}z = \mathbf{b}$ is easily solved and all eigenvalues of $\mathbf{M}^{-1}\mathbf{N}$ lie on the real subinterval $[\alpha, \beta]$ of $[-1, 1]$:

1. Let $\gamma := \dfrac{2}{2-\alpha-\beta}$ and $\mu := \dfrac{2-\alpha-\beta}{\beta-\alpha}$;
2. Take an arbitrary starting vector $\mathbf{y}_0 := \mathbf{x}_0$;
   Take $\omega_1 := 2$;
   Solve $\mathbf{M}\mathbf{z}_0 = \mathbf{b} - \mathbf{A}\mathbf{y}_0$ for $\mathbf{z}_0$;
   Let $\mathbf{y}_1 := \mathbf{x}_0 + \gamma\mathbf{z}_0$ (3.42);

© 2003 by CRC Press LLC

3. For $k = 1, 2, \ldots$:

Let $\omega_{k+1} := \dfrac{1}{1 - \omega_k / 4\mu^2}$ (3.44);

Solve $\mathbf{M}\mathbf{z}_k = \mathbf{b} - \mathbf{A}\mathbf{y}_k$ for $\mathbf{z}_k$ (3.43);

Let $\mathbf{y}_{k+1} := \omega_{k+1}(\mathbf{y}_k - \mathbf{y}_{k-1} + \gamma\mathbf{z}_k) + \mathbf{y}_{k-1}$ (3.41).

3.5 Telescoping procedures for power series

If a function $f(x)$ may be expanded in a power series which converges on $[-1, 1]$ (possibly after a suitable transformation of the x variable), then a plausible approximation may clearly be obtained by truncating this power series after $n+1$ terms to a polynomial $p_n(x)$ of degree n. It may be possible, however, to construct an nth degree polynomial approximation better than this, by first truncating the series to a polynomial $p_m(x)$ of some higher degree $m > n$ (which will usually be a better approximation to $f(x)$ than $p_n(x)$) and then exploiting the properties of Chebyshev polynomials to 'economise' $p_m(x)$ to a polynomial of degree n.

The simplest economisation technique is based on the idea of subtracting a constant multiple of a Chebyshev polynomial of the same degree, the constant being chosen so as to reduce the degree of the polynomial.

EXAMPLE 3.3: For $f(x) = \mathrm{e}^x$, the partial sum of degree 7 of the power series expansion is given by

$$\begin{aligned} p_7(x) &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^7}{7!} \\ &= 1 + x + 0.5x^2 + 0.1666667x^3 + 0.0416667x^4 + \\ &\qquad + 0.008333x^5 + 0.0013889x^6 + 0.0001984x^7, \end{aligned} \tag{3.45}$$

where a bound on the error in approximating $f(x)$ is given, by the mean value theorem, by

$$|f(x) - p_7(x)| = \left|\frac{x^8}{8!} f^{(8)}(\xi)\right| = \left|\frac{x^8}{8!}\mathrm{e}^{\xi}\right| \le \frac{\mathrm{e}}{8!} = 0.0000674 \text{ for } x \text{ in } [-1, 1]. \tag{3.46}$$

(The actual maximum error on $[-1, 1]$ in this example is in fact the error at $x = 1$, $|f(1) - p_7(1)| = 0.0000279$.)

Now (3.45) may be economised by forming the degree-6 polynomial

$$\begin{aligned} p_6(x) &= p_7(x) - 0.0001984\ [2^{-6}T_7(x)] \\ &= p_7(x) - 0.0000031\ T_7(x). \end{aligned} \tag{3.47}$$

Since $2^{-6}T_7(x)$ is the minimax monic polynomial of degree 7, this means that p_6 is the minimax 6th degree approximation to p_7 on $[-1, 1]$, and p_7 has been economised in an optimal way.

© 2003 by CRC Press LLC

From (3.45), (3.47) and the coefficients in Table C.2, we obtain

$$\begin{aligned} p_6(x) &= p_7(x) - 0.0001984(64x^7 - 112x^5 + 56x^3 - 7x)/2^6 \\ &= p_7(x) - 0.0001984(x^7 - 1.75x^5 + 0.875x^3 - 0.109375x). \end{aligned}$$

Thus

$$\begin{aligned} p_6(x) &= 1 + 1.0000217x + 0.5x^3 + 0.1664931x^3 + \\ &\quad + 0.0416667x^4 + 0.0086805x^5 + 0.0013889x^6. \end{aligned} \tag{3.48}$$

(Since $T_7(x)$ is an odd function of x, coefficients of even powers of x are unchanged from those in $p_7(x)$.) An error has been committed in replacing p_7 by p_6, and, from (3.47), this error is of magnitude 0.0000031 at most (since $|T_7(x)|$ is bounded by 1 on the interval). Hence, from (3.46), the accumulated error in $f(x)$ satisfies

$$|f(x) - p_6(x)| \le 0.0000674 + 0.0000031 = 0.0000705. \tag{3.49}$$

A further economisation leads to the quintic polynomial

$$\begin{aligned} p_5(x) &= p_6(x) - 0.0013889\ [2^{-5}T_6(x)] \\ &= p_6(x) - 0.0000434\, T_6(x). \end{aligned} \tag{3.50}$$

Here p_5 is the minimax quintic polynomial approximation to p_6. From (3.48), (3.50) and Table C.2, we obtain

$$\begin{aligned} p_5(x) &= p_6(x) - 0.0013889(32x^6 - 48x^2 + 18x^2 - 1)/2^5 \\ &= p_6(x) - 0.0013889(x^6 - 1.5x^4 + 0.5625x^2 - 0.03125). \end{aligned}$$

Thus

$$\begin{aligned} p_5(x) &= 1.0000062 + 1.0000217x + 0.4992188x^2 + \\ &\quad + 0.1664931x^3 + 0.0437500x^4 + 0.0086805x^5 \end{aligned} \tag{3.51}$$

and, since $T_6(x)$ is an even function of x, coefficients of odd powers are unchanged from those in $p_6(x)$. The error in replacing p_6 by p_5 is, from (3.50), at most 0.0000434. Hence, from (3.49), the accumulated error in $f(x)$ now satisfies

$$|f(x) - p_5(x)| \le 0.0000705 + 0.0000434 = 0.0001139. \tag{3.52}$$

Thus the degradation in replacing p_7 (3.45) by p_5 (3.51) is only marginal, increasing the error bound from 0.000067 to 0.000114.

In contrast, the partial sum of degree 5 of the power series (3.45) has a mean-value-theorem error bound of $\left|x^6 \mathrm{e}^\xi/6!\right| \le \mathrm{e}/6! \sim 0.0038$ on $[-1, 1]$, and the actual maximum error on $[-1, 1]$, attained at $x = 1$, is 0.0016. However, even this is about 15 times as large as (3.52), so that the telescoping procedure based on Chebyshev polynomials is seen to give a greatly superior approximation.

The approximation (3.50) and the 5th degree partial sum of the Taylor series are both too close to e^x for the error to be conveniently shown graphically. However, in Figures 3.3 and 3.4 we show the corresponding approximations of degree 2, where the improved accuracy is clearly visible.

© 2003 by CRC Press LLC

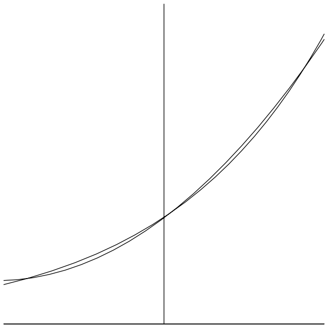

Figure 3.3: The function e^x on $[-1, 1]$ and an economised polynomial approximation of degree 2

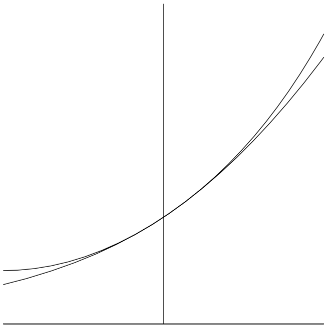

Figure 3.4: The function e^x on $[-1, 1]$ and its Taylor series truncated at the 2nd degree term

An alternative technique which might occur to the reader is to rewrite the polynomial $p_7(x)$, given by (3.45), as a sum of Chebyshev polynomials

$$p_7(x) = \sum_{k=0}^{7} c_k T_k(x), \tag{3.53}$$

where c_k are determined by using, for example, the algorithm of Section 2.3.1 above (powers of x in terms of $\{T_k(x)\}$). Suitable higher order terms, such as those in T_6 and T_7, could then be left out of (3.53) according to the size of their coefficients c_k. However, the telescoping procedure above is exactly equivalent to this, and is in fact a somewhat simpler way of carrying it out. Indeed c_7 and c_6 have been calculated above, in (3.47) and (3.50) respectively, as

$$c_7 = 0.0000031, \quad c_6 = 0.0000434.$$

If the telescoping procedure is continued until a constant approximation $p_0(x)$ is obtained, then all of the Chebyshev polynomial coefficients c_k will be determined.

3.5.1 Shifted Chebyshev polynomials on $[0, 1]$

The telescoping procedure may be adapted to ranges other than $[-1, 1]$, provided that the Chebyshev polynomials are adjusted to the range required. For example, the range $[-c, c]$ involves the use of the polynomials $T_k(x/c)$. A range that is often useful is $[0, 1]$ (or, by scaling, $[0, c]$), and in that case the shifted Chebyshev polynomials $T_k^*(x)$ (or $T_k^*(x/c)$) are used. Since the latter polynomials are neither even nor odd, every surviving coefficient in the polynomial approximation changes at each economisation step.

© 2003 by CRC Press LLC

EXAMPLE 3.4: Suppose that we wish to economise on $[0, 1]$ a quartic approximation to $f(x) = \mathrm{e}^x$:

$$q_4(x) = 1 + x + 0.5x^2 + 0.1666667x^3 + 0.0416667x^4$$

in which the error satisfies

$$|f(x) - q_4(x)| = \frac{x^5}{5!}\mathrm{e}^{\xi} \le \frac{\mathrm{e}}{5!} = 0.0227. \tag{3.54}$$

Then the first economisation step leads to

$$\begin{aligned} q_3(x) &= q_4(x) - 0.0416667\ [2^{-7}T_4^*(x)] \\ &= q_4(x) - 0.0003255\ T_4^*(x). \end{aligned} \tag{3.55}$$

From Table C.2:

$$\begin{aligned} q_3(x) &= q_4(x) - 0.0416667(128x^4 - 256x^3 + 160x^2 - 32x + 1)/2^7 \\ &= q_4(x) - 0.0416667(x^4 - 2x^3 + 1.25x^2 - 0.25x + 0.0078125). \end{aligned}$$

Thus

$$q_3(x) = 0.9996745 + 1.0104166x + 0.4479167x^2 + 0.25x^3. \tag{3.56}$$

Here the maximum additional error due to the economisation is 0.0003255, from (3.55), which is virtually negligible compared with the existing error (3.54) of q_4. In fact, the maximum error of (3.56) on $[0, 1]$ is 0.0103, whereas the maximum error of the power series truncated after the term in x^3 is 0.0516.

The economisation can be continued in a similar way for as many steps as are valid and necessary. It is clear that significantly smaller errors are incurred on $[0, 1]$ by using $T_k^*(x)$ than are incurred on $[-1, 1]$ using $T_k(x)$. This is to be expected, since the range is smaller. Indeed there is always a reduction in error by a factor of 2^m, in economising a polynomial of degree m, since the respective monic polynomials that are adopted are

$$2^{1-m}T_m(x) \text{ and } 2^{1-2m}T_m^*(x).$$

3.5.2 Implementation of efficient algorithms

The telescoping procedures above, based on $T_k(x)$ and $T_k^*(x)$ respectively, are more efficiently carried out in practice by implicitly including the computation of the coefficients of the powers of x in T_k or T_k^* within the procedure (so that Table C.2 does not need to be stored). This is best achieved by using ratios of consecutive coefficients from formula (2.19) of Section 2.3.3 above.

© 2003 by CRC Press LLC

Consider first the use of the shifted polynomial $T_k^*(x/d)$ on a chosen range $[0, d]$. Suppose that $f(x)$ is initially approximated by a polynomial $p_m(x)$ of degree m, where for each $\ell \le m$,

$$p_\ell(x) = \sum_{k=0}^{\ell} a_k^{(\ell)} x^k = a_0^{(\ell)} + a_1^{(\ell)} x + \cdots + a_\ell^{(\ell)} x^\ell. \tag{3.57}$$

Then the first step of the telescoping procedure replaces p_m by

$$p_{m-1}(x) = p_m(x) - a_m^{(m)} 2^{1-2m} d^m T_m^*(x/d). \tag{3.58}$$

(The factor d^m is included, to ensure that $2^{1-2m} d^m T_m^*(x/d)$ is monic.)

Now, write

$$T_m^*(x/d) = 2^{2m-1} d^{-m} \sum d_k^{(m)} x^k d^{m-k} \tag{3.59}$$

where $2^{2m-1} d_k^{(m)}$ is the coefficient of x^k in $T_m^*(x)$. Then, by (3.57), (3.58), (3.59):

$$a_k^{(m-1)} = a_k^{(m)} - a_m^{(m)} d_k^{(m)} \quad (k = m-1, m-2, \ldots, 0). \tag{3.60}$$

The index k has been ordered from $k = m-1$ to $k = 0$ in (3.60), since the coefficients $d_k^{(m)}$ will be calculated in reverse order below.

Now $T_m^*(x) = T_{2m}(x^{\frac{1}{2}})$ and hence, from (2.16),

$$T_m^*(x) = \sum_{k=0}^{m} c_k^{(2m)} x^{m-k} \tag{3.61}$$

where $c_k^{(2m)}$ is defined by (2.17a). Hence, in (3.59),

$$d_k^{(m)} = c_{m-k}^{(2m)} 2^{1-2m}. \tag{3.62}$$

Now, from (2.19)

$$c_{k+1}^{(n)} = -\frac{(n-2k)(n-2k-1)}{4(k+1)(n-k-1)} c_k^{(n)} \tag{3.63}$$

and hence, from (3.62),

$$d_{m-k-1}^{(m)} = -\frac{(2m-2k)(2m-2k-1)}{4(k+1)(2m-k-1)} d_{m-k}^{(m)}.$$

Thus

$$d_{r-1}^{(m)} = \frac{-2r(2r-1)}{4(m-r+1)(m+r-1)} d_r^{(m)} \tag{3.64}$$

where $d_m^{(m)} = 1$.

In summary, the algorithm is as follows:

© 2003 by CRC Press LLC

Given $p_m(x)$ of form (3.57), with coefficients $a_k^{(m)}$:

1. With $d_m^{(m)} = 1$, determine $d_{m-1}^{(m)}, \ldots, d_0^{(m)}$, using (3.64);
2. Determine $a_k^{(m-1)}$, using (3.60), and hence $p_{m-1}(x)$ of form (3.57), with coefficients $a_k^{(m-1)}$.

However, if a telescoping procedure is based on the range $[-d, d]$ and the standard polynomials $T_k(x/d)$, then it is more appropriate to treat even and odd powers of x separately, since each T_k involves only one or the other, and so the algorithm is correspondingly more complicated, but at the same time more efficient.

Suppose $f(x)$ is initially approximated by the polynomial $p_{2M+1}(x)$ of odd degree, where (for each $\ell \le M$)

$$p_{2\ell+1}(x) = \sum_{k=0}^{\ell} b_k^{(\ell)} x^{2k+1} + \sum_{k=0}^{\ell} c_k^{(\ell)} x^{2k} \tag{3.65a}$$

and

$$p_{2\ell}(x) = \sum_{k=0}^{\ell-1} b_k^{(\ell-1)} x^{2k+1} + \sum_{k=0}^{\ell} c_k^{(\ell)} x^{2k}. \tag{3.65b}$$

Then the first two (odd and even) steps of the telescoping procedure replace $p_{2M+1}(x)$ by $p_{2M}(x)$ and $p_{2M}(x)$ by $p_{2M-1}(x)$, where

$$p_{2M}(x) = p_{2M+1}(x) - b_M^{(M)} 2^{-2M} d^{2M+1} T_{2M+1}(x/d), \tag{3.66a}$$

$$p_{2M-1}(x) = p_{2M}(x) - c_M^{(M)} 2^{1-2M} d^{2M} T_{2M}(x/d). \tag{3.66b}$$

Now, let $2^{2M} e_k^{(M)}$ and $2^{2M-1} f_k^{(M)}$ denote respectively the coefficients of x^{2k+1} in $T_{2M+1}(x)$ and of x^{2k} in $T_{2M}(x)$.

Then, from (2.16),

$$T_{2M+1}(x/d) = 2^{2M} d^{-2M-1} \sum_{k=0}^{M} e_k^{(M)} x^{2k+1} d^{2M-2k} =$$

$$= \sum_{k=0}^{M} b_{M-k}^{(M)} x^{2k+1} d^{-2k-1} \tag{3.67a}$$

$$T_{2M}(x/d) = 2^{2M-1} d^{-2M} \sum_{k=0}^{M} f_k^{(M)} x^{2k} d^{2M-2k} =$$

$$= \sum_{k=0}^{M} c_{M-k}^{(M)} x^{2k} d^{-2k} \tag{3.67b}$$

© 2003 by CRC Press LLC

Hence, from (3.65)–(3.67),

$$b_k^{(M-1)} = b_k^{(M)} - b_M^{(M)} e_k^{(M)} \quad (k = M-1, M-2, \ldots, 0), \tag{3.68a}$$

$$c_k^{(M-1)} = c_k^{(M)} - c_M^{(M)} f_k^{(M)} \quad (k = M-1, M-2, \ldots, 0). \tag{3.68b}$$

Formulae for generating the scaled Chebyshev coefficients $e_k^{(M)}$ and $f_k^{(M)}$ may be determined from (3.63) and (3.67) (by replacing n by $2M+1$, $2M$, respectively) in the form

$$\begin{aligned} e_{M-k-1}^{(M)} &= -\frac{(2M-2k+1)(2M-2k)}{4(k+1)(2M-k)} e_{M-k}^{(M)}, \\ f_{M-k-1}^{(M)} &= -\frac{(2M-2k)(2M-2k-1)}{4(k+1)(2M-k-1)} f_{M-k}^{(M)}. \end{aligned}$$

Thus $e_M^{(M)} = f_M^{(M)} = 1$, and

$$e_{r-1}^{(M)} = -\frac{(2r+1)(2r)}{4(M-r+1)(M+r)} e_r^{(M)}, \tag{3.69a}$$

$$f_{r-1}^{(M)} = -\frac{(2r)(2r-1)}{4(M-r+1)(M+r-1)} f_r^{(M)}. \tag{3.69b}$$

In summary, the algorithm is as follows:

Given $p_{2M+1}(x)$ of form (3.65a), with coefficients $b_k^{(M)}$ and $c_k^{(M)}$:

1. With $e_M^{(M)} = 1$, determine $e_{M-1}^{(M)}, \ldots, e_0^{(M)}$, using (3.69a);
2. Determine $b_k^{(M-1)}$, using (3.68a), and hence $p_{2M}(x)$ of form (3.65b), with coefficients $b_k^{(M-1)}$ and $c_k^{(M)}$;
3. With $f_M^{(M)} = 1$, determine $f_{M-1}^{(M)}, \ldots, f_0^{(M)}$, using (3.69b);
4. Determine $c_k^{(M-1)}$, using (3.68b), and hence $p_{2M-1}(x)$ of form (3.65a), with coefficients $b_k^{(M-1)}$ and $c_k^{(M-1)}$.

We should add as a postscript that Gutknecht & Trefethen (1982) have succeeded in implementing an alternative economisation method due to Carathéodory and Fejér, which yields a Chebyshev sum giving a much closer approximation to the original polynomial.

© 2003 by CRC Press LLC

3.6 The tau method for series and rational functions

Sometimes a power series converges very slowly at a point of interest, or even diverges, so that we cannot find a suitable partial sum to provide an initial approximation for the above telescoping procedure. However, in some cases other approaches are useful, one of which is the 'tau' (τ) method[1] of Lanczos (1957).

Consider for example the function

$$y(x) = \frac{1}{1+x}$$

which has the power series expansion

$$1 - x + x^2 - x^3 + \cdots + (-1)^n x^n + \cdots.$$

This series has radius of convergence 1, and since it does not converge for $|x| \geq 1$, cannot be used on $[0, 1]$ or wider ranges. However, $y(x)$ is the solution of the functional equation

$$(1 + x)y(x) = 1 \tag{3.70}$$

and may be approximated on $[0, 1]$ by a polynomial $p_n(x)$ of degree n in the form

$$p_n(x) = {\sum_{k=0}^{n}}' c_k T_k^*(x) \tag{3.71}$$

(where, as previously, the dash denotes that the first term in the sum is halved), by choosing the coefficients $\{c_k\}$ so that p_n approximately satisfies the equation

$$(1 + x)p_n(x) = 1. \tag{3.72}$$

Equation (3.72) can be perturbed slightly into one that can be satisfied exactly, by adding to the right-hand side an undetermined multiple τ (say) of a shifted Chebyshev polynomial of degree $n + 1$:

$$(1 + x)p_n(x) = 1 + \tau T_{n+1}^*(x). \tag{3.73}$$

Since there are $n + 2$ free parameters in (3.71) and (3.73), namely c_k ($k = 0, 1, \ldots, n$) and τ, it should be possible to determine them by equating coefficients of powers of x in (3.73) (since there are $n+2$ coefficients in a polynomial of degree $n+1$). Equivalently, we may equate coefficients of Chebyshev polynomials after writing the two sides of (3.73) as Chebyshev summations; this can be done if we note from (2.39) that

$$(2x - 1)T_k(2x - 1) = \tfrac{1}{2}[T_{k+1}(2x - 1) + T_{|k-1|}(2x - 1)]$$

[1] A slightly different but related approach, also known as the 'tau method', is applied to solve differential equations in a later chapter (see Chapter 10).

© 2003 by CRC Press LLC

and hence, since $T_k^*(x) = T_k(2x-1)$,

$$(1+x)T_k^*(x) = \tfrac{1}{4}[T_{k+1}^*(x) + 6T_k^*(x) + T_{|k-1|}^*(x)]. \tag{3.74}$$

Substituting (3.74) into (3.71) and (3.73),

$$\sum_{k=0}^{n}{}' \tfrac{1}{4}c_k[T_{|k-1|}^*(x) + 6T_k^*(x) + T_{k+1}^*(x)] = T_0^*(x) + \tau T_{n+1}^*(x).$$

On equating coefficients of $T_0^*, \ldots, T_{n+1}^*$, we obtain

$$\begin{aligned}
\tfrac{1}{4}(3c_0 + c_1) &= 1, \\
\tfrac{1}{4}(c_{k-1} + 6c_k + c_{k+1}) &= 0 \quad (k = 1, \ldots, n-1), \\
\tfrac{1}{4}(c_{n-1} + 6c_n) &= 0, \\
\tfrac{1}{4}c_n &= \tau.
\end{aligned}$$

These are $n+2$ equations for c_0, c_1, $\ldots$, c_n and τ, which may be readily solved by back-substituting for c_n in terms of τ, hence (working backwards) determining c_{n-1} , c_{n-2}, $\ldots$, c_0 in terms of τ, leaving the first equation to determine the value of τ.

EXAMPLE 3.5: For $n = 3$, we obtain (in this order)

$$\begin{aligned}
c_3 &= 4\tau, \\
c_2 &= -6c_3 = -24\tau, \\
c_1 &= -6c_2 - c_3 = 140\tau, \\
c_0 &= -6c_1 - c_2 = -816\tau, \\
3c_0 + c_1 &= -2308\tau = 4.
\end{aligned}$$

Hence $\tau = -1/577$ and, from (3.71),

$$\begin{aligned}
y(x) &\simeq p_3(x) \\
&= \tfrac{1}{577}[408T_0^*(x) - 140T_1^*(x) + 24T_2^*(x) - 4T_3^*(x)] \\
&= 0.707106T_0^*(x) - 0.242634T_1^*(x) + \\
&\quad + 0.041594T_2^*(x) - 0.006932T_3^*(x).
\end{aligned} \tag{3.75}$$

The error $\epsilon(x)$ in (3.75) is known from (3.70) and (3.73) to be

$$\epsilon(x) = y(x) - p_3(x) = \frac{\tau T_4^*(x)}{1+x}.$$

Since $1/(1+x)$ and $T_4^*(x)$ are both bounded by 1 in magnitude, we deduce the bound

$$|\epsilon(x)| \le |\tau| = 0.001704 \simeq 0.002 \text{ on } [0,1]. \tag{3.76}$$

This upper bound is attained at $x = 0$, and we would expect the resulting approximation $p_3(x)$ to be reasonably close to a minimax approximation.

© 2003 by CRC Press LLC

3.6.1 The extended tau method

Essentially the same approach has been proposed by Fox & Parker (1968) for the approximation on $[-1, 1]$ of a rational function $a(x)/b(x)$ of degrees (p, q). They introduce a perturbation polynomial

$$e(x) = \sum_{m=n+1}^{n+q} \tau_{m-n} T_m(x), \tag{3.77}$$

in place of the single term $\tau T_{n+1}(x)$ used above, to give

$$a(x) + e(x) = b(x) \sum_{k=0}^{n}{}' c_k T_k(x). \tag{3.78}$$

The number and degrees of the terms in (3.77) are chosen so that (3.78) is uniquely solvable for $\{c_k\}$ and $\{\tau_m\}$.

For example, for

$$\frac{a(x)}{b(x)} = \frac{1 - x + x^2}{1 + x + x^2}$$

we need two tau terms and (3.78) becomes

$$(1 - x + x^2) + \sum_{m=n+1}^{n+2} \tau_{m-n} T_m(x) = (1 + x + x^2) \sum_{k=0}^{n}{}' c_k T_k(x). \tag{3.79}$$

Both sides of (3.79) are then written in terms of Chebyshev polynomials, and on equating coefficients, a set of equations is obtained for c_k and τ_m. Back-substitution in terms of τ_1 and τ_2 leads to a pair of simultaneous equations for τ_1 and τ_2; hence c_k are found.

Example 3.6: For $n = 2$, (3.79) becomes, using (2.38) to transform products into sums,

$$\begin{aligned}
(\tfrac{3}{2}T_0(x) \;&-\; T_1(x) + \tfrac{1}{2}T_2(x)) + \tau_1 T_3(x) + \tau_2 T_4(x) \\
&= (\tfrac{3}{2}T_0(x) + T_1(x) + \tfrac{1}{2}T_2(x))(\tfrac{1}{2}c_0 T_0(x) + c_1 T_1(x) + c_2 T_2(x)) \\
&= (\tfrac{3}{4}c_0 + \tfrac{1}{2}c_1 + \tfrac{1}{4}c_2)T_0(x) + (\tfrac{1}{2}c_0 + \tfrac{7}{4}c_1)T_1(x) + \\
&\qquad + (\tfrac{1}{4}c_0 + \tfrac{1}{2}c_1 + \tfrac{3}{2}c_2)T_2(x) + (\tfrac{1}{4}c_1 + \tfrac{1}{2}c_2)T_3(x) + \tfrac{1}{4}T_4(x).
\end{aligned}$$

Equating coefficients of the Chebyshev polynomials $T_0(x)$, ..., $T_4(x)$ yields the equations

$$\left.\begin{aligned}
c_2 &= 4\tau_2 \\
c_1 + 2c_2 &= 4\tau_1 \\
c_0 + 2c_1 + 6c_2 &= 2
\end{aligned}\right\} \tag{3.80}$$

$$\left.\begin{aligned}
2c_0 + 7c_1 &= -4 \\
3c_0 + 2c_1 + c_2 &= 6
\end{aligned}\right\} \tag{3.81}$$

© 2003 by CRC Press LLC

Back-substituting in (3.80):

$$c_2 = 4\tau_2,\ c_1 = 4\tau_1 - 8\tau_2,\ c_0 = 2 - 8\tau_1 - 8\tau_2.$$

Now (3.81) gives

$$\begin{aligned} 3\tau_1 - 16\tau_2 &= -2 \\ 9\tau_1 + 4\tau_2 &= 0 \end{aligned}$$

and hence

$$\tau_1 = -18/91,\ \tau_2 = 8/91$$

so that

$$c_0 = 262/91,\ c_1 = -136/91,\ c_2 = 32/91.$$

Thus

$$y(x) = \frac{a(x)}{b(x)} = \frac{1 - x + x^2}{1 + x + x^2} \simeq p_3(x) = 1.439\, T_0(x) - 1.494\, T_1(x) + 0.352\, T_2(x)$$

and the error is given by

$$\begin{aligned} \epsilon(x) = y(x) - p_3(x) &= -\frac{\tau_1 T_3(x) + \tau_2 T_4(x)}{1 + x + x^2} \\ &= \frac{0.198\, T_3(x) - 0.088\, T_4(x)}{1 + x + x^2}. \end{aligned}$$

On $[-1, 1]$, $1/(1 + x + x^2)$ is bounded by $\frac{4}{3}$ and $|T_3|$ and $|T_4|$ are bounded by 1. Hence we have the bound (which is not far from the actual maximum error)

$$|\epsilon(x)| < 1.333\ (0.198 + 0.088) = 0.381.$$

With an error bound of 0.381, the approximation found in this example is not particularly accurate, and indeed a much higher degree of polynomial is needed to represent such a rational function at all reasonably, but the method does give credible and measurable results even in this simple case (see Figure 3.5).

We may note that an alternative approach to the whole calculation is to use the power form for the polynomial approximation

$$p_n(x) = \sum_{k=0}^{n} a_k x^k \tag{3.82}$$

and then to replace (3.79) by

$$(1 - x + x^2) + \sum_{m=n+1}^{n+2} \tau_{m-n} T_m(x) = (1 + x + x^2) \sum_{k=0}^{n} a_k x^k. \tag{3.83}$$

We then equate coefficients of powers of x and solve for τ_1 and τ_2.

© 2003 by CRC Press LLC

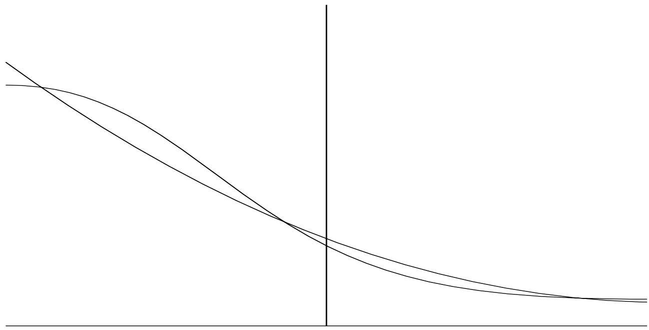

Figure 3.5: Rational function and a quadratic approximation obtained by the extended τ method

---

EXAMPLE 3.7: For $n = 2$, equation (3.83) takes the form

$$(1 - x + x^2) + \tau_1(4x^3 - 3x) + \tau_2(8x^4 - 8x^2 + 1) = (1 + x + x^2)(a_0 + a_1 x + a_2 x^2),$$

and on equating coefficients of $1, x, \ldots, x^4$,

$$\left.\begin{array}{rcl} a_0 &=& 1 + \tau_2 \\ a_1 + a_0 &=& -1 - 3\tau_1 \end{array}\right\} \tag{3.84}$$

$$\left.\begin{array}{lcl} a_2 + a_1 + a_0 &=& 1 - 8\tau_2 \\ a_2 + a_1 &=& 4\tau_1 \\ a_2 &=& 8\tau_2 \end{array}\right\} \tag{3.85}$$

Back-substituting in (3.85):

$$a_2 = 8\tau_2, \;\; a_1 = 4\tau_1 - 8\tau_2, \;\; a_0 = 1 - 4\tau_1 - 8\tau_2.$$

Now (3.84) gives

$$4\tau_1 + 9\tau_2 \;=\; 0$$

$$3\tau_1 - 16\tau_2 \;=\; -2$$

and hence

$$\tau_1 = -18/91, \;\; \tau_2 = 8/91$$

(the same values as before) so that

$$a_0 = 99/91, \;\; a_1 = -136/91, \;\; a_2 = 64/91.$$

Thus

$$y(x) = \frac{a(x)}{b(x)} = \frac{1 - x + x^2}{1 + x + x^2} \simeq p_3(x) = 1.088 - 1.495x + 0.703x^2.$$

It is easily verified that this is precisely the same approximation as was obtained previously, but expressed explicitly as a sum of powers of x.

© 2003 by CRC Press LLC

For the degrees n of polynomial likely to be required in practice, it is not advisable to use the power representation (3.82), even though the algebra appears simpler, since the coefficients a_k tend to become large as n increases, whereas the Chebyshev coefficients c_k in the form (3.71) typically tend to converge with n to the true coefficients of an infinite Chebyshev series expansion (see Chapter 4).

3.7 Problems for Chapter 3

1. Verify the axioms of a vector space for the following families of functions or data:

   (a) $\mathcal{F} = \mathcal{C}[a, b]$;

   (b) $\mathcal{F} = \{\{f(x_k),\ k = 1, \ldots, m\}\}$ (values of a function at discrete points).

   What are the dimensions of these spaces?

2. Verify, from the definition of a norm, that the following is a norm:

$$\|f\| = \|f\|_p = \left[\int_a^b |f(x)|^p \,\mathrm{d}x\right]^{\frac{1}{p}} \qquad (1 \le p < \infty),$$

   by assuming Minkowski's continuous inequality:

$$\left(\int |f + g|^p \,\mathrm{d}x\right)^{\frac{1}{p}} \le \left(\int |f|^p \,\mathrm{d}x\right)^{\frac{1}{p}} + \left(\int |g|^p \,\mathrm{d}x\right)^{\frac{1}{p}}.$$

   Prove the latter inequality for $p = 1, 2$, and show, for $p = 2$, that equality does not occur unless $f(x) = \lambda g(x)$ ('almost everywhere'), where λ is some constant.

3. For what values of p does the function $f(x) = (1 - x^2)^{-1/2}$ belong to the function space $\mathcal{L}_p[-1, 1]$, and what is its norm?

4. Prove Minkowski's discrete inequality:

$$\left(\sum_k |u_k + v_k|^p\right)^{\frac{1}{p}} \le \left(\sum_k |u_k|^p\right)^{\frac{1}{p}} + \left(\sum_k |v_k|^p\right)^{\frac{1}{p}}$$

   in the case $p = 2$ by first showing that

$$\left(\sum u_k v_k\right)^2 \le \sum (u_k)^2 \sum (v_k)^2.$$

© 2003 by CRC Press LLC

Deduce that

$$\|f\|_p = \left[\sum_{k=1}^{m} |f(x_k)|^p \right]^{\frac{1}{p}} \qquad (1 \le p < \infty)$$

is a norm for space (b) of Problem 1.

Find proofs in the literature (Hardy et al. 1952, for example) of both continuous and discrete Minkowski inequalities for general p. Can equality occur for $p = 1$?

5. Find the minimax constant (i.e., polynomial of degree zero) approximation to e^x on $[-1, 1]$, by assuming that its error has the alternating property at $-1, +1$. Deduce that the minimax error in this case is $\sinh 1$.

   Generalise the above approach to determine a minimax constant approximation to any monotonic continuous function $f(x)$.

6. Prove the *sufficiency* of the characterisation of the error in Theorem 3.4, namely that, for a polynomial approximation p_n of degree n to a continuous f to be minimax, it is sufficient that it should have the alternating property at $n + 2$ points $x_1 < \cdots < x_{n+2}$.

   [Hint: Assume that an approximation p_n' exists with smaller error norm than p_n, show that $p_n - p_n'$ changes sign between each pair x_i and x_{i+1}, and hence obtain the result.]

7. Consider the function

$$f(x) = \sum_{i=0}^{\infty} c_i T_{b^i}(x), \qquad (*)$$

   where $\{c_i\}$ are so defined that the series is uniformly convergent and where b is an odd integer not less than 2. Show that, for every $i > n$ with n fixed, T_{b^i} has the alternating property on a set of $b^n + 1$ consecutive points of $[-1, 1]$. Deduce that the partial sum of degree b^n of (*) (namely the sum from $i = 0$ to n) is the minimax polynomial approximation of degree b^n to $f(x)$.

   [Note: A series in $\{T_k(x)\}$ such as (*) in which terms occur progressively more rarely (in this case for $k = 0, b, b^2, b^3, \ldots$) is called *lacunary*; see Section 5.9 below for a fuller discussion.]

8. For $f(x) = \arctan x$, show that $(1 + x^2)f'(x) = 1$, and hence that

$$(1+x^2)f^{(n)}(x)+2x(n-1)f^{(n-1)}(x)+(n+1)(n+2)f^{(n-2)}(x) = 0 \quad (n \ge 2).$$

   Deduce the Taylor–Maclaurin expansion

$$f(x) \sim x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} + \cdots. \qquad (**)$$

© 2003 by CRC Press LLC

Estimate the error in the partial sum $P_7(x)$ of degree 7 of (**) for x in $[-0.3, 0.3]$.

Telescope P_7, into polynomials P_5 of degree 5 and P_3 of degree 3 by using Chebyshev polynomials normalised to $[-0.3, 0.3]$, and estimate the accumulated errors in P_5 and P_3.

9. Given

$$f(x) = \log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^{n-1}\frac{x^n}{n} + \cdots, \qquad (***)$$

use the mean value theorem to give a bound on the error on $[0, 0.1]$ of the partial sum P_n of degree n of (***). Telescope P_4 into polynomials P_3 of degree 3 and P_2 of degree 2, respectively, using a Chebyshev polynomial adjusted to $[0, 0.1]$, and estimate the accumulated errors in each case.

10. (Programming Exercise) Write a computer program (in a programming language of your own choice) to implement the telescoping algorithm of Section 3.5, either

    (a) based on $T_k^*(x/d)$ and using (3.60)–(3.64) or

    (b) based on $T_k(x/d)$ and using (3.68)–(3.69).

11. Apply the tau method of Section 3.6 to determine a polynomial approximation of degree 3 to $x/(1+x)$ on $[0, 1]$ based on the equation

$$(1+x)y = x$$

and determine a bound on the resulting error.

12. Apply the extended tau method of Section 3.6.1 to determine a polynomial approximation of degree 2 to $(1+x+x^2)^{-1}$ on $[-1, 1]$ and determine a bound on the resulting error.

13. Show that $2^{-n}\sqrt{1-x^2}U_n(x)$, $2^{-n}\sqrt{1+x}V_n(x)$ and $2^{-n}\sqrt{1-x}W_n(x)$ equioscillate on $(n+2)$, $(n+1)$ and $(n+1)$ points, respectively, of $[-1, 1]$, and find the positions of their extrema. Deduce that these are minimax approximations to zero by monic polynomials of degree n with respective weight functions $\sqrt{1-x^2}$, $\sqrt{1+x}$, $\sqrt{1-x}$. Why are there more equioscillation points in the first case?

© 2003 by CRC Press LLC