JOURNAL OF APPROXIMATION THEORY 17, 254–279 (1976)

The Uniform Approximation of Polynomials by Polynomials of Lower Degree

A. Talbot

Mathematics Department, Brunel University, Uxbridge, Middlesex, England

Communicated by C. W. Clenshaw

Received October 14, 1974

1. Introduction

The general problem of finding the best uniform approximation, in a given interval, of a polynomial of degree m by a polynomial of degree $n < m$ has been solved analytically in only two cases: (i) by Chebyshev, when $m = n + 1$, (ii) by Zolotarev, when $m = n + 2$. In case (i) the solution is expressible in terms of the Chebyshev polynomial $T_m(x)$. In case (ii) the solution (see for example Achieser [1, p. 280]) involves elliptic functions. Chebyshev did in fact consider the general case in [4], and showed that hyperelliptic functions are involved, but he did not obtain any solutions.

Since analytic solutions are effectively excluded when $m > n + 2$, another approach is required. This was first provided, for large n, by Bernstein [3] and Achieser [2]. It consists in seeking a rational function which (a) is a good approximation to the given polynomial, and (b) has a fractional part which for large n is small in the interval. Its integral part is then the polynomial approximation desired: not optimal, but asymptotically optimal.

In 1964 Clenshaw [6] considered the ratio S_n/E_n of the uniform error norms S_n and E_n, respectively, of the truncated Chebyshev expansion of the given polynomial and the best uniform approximation. He used Bernstein's method to estimate E_n when $m - n = 2$, 3, or 4, but could go no further because of the complication of the calculations. Clenshaw was interested in a question of practical importance, namely, whether the truncated Chebyshev expansion, which is easy to obtain, is or is not nearly as good an approximation as the optimum. He therefore tackled the problem of finding the maximum value of S_n/E_n for a given $m - n$. Subject to an assumption which he verified experimentally, Clenshaw solved the problem for the three cases mentioned, and noticed some surprising regularities in the solution, in particular the fact that certain constants obtained were the first 2, 3, and 4 coefficients, respectively, of the binomial expansion of $(1-t)^{-1/2}$. He put forward the conjecture that this would generalize for any value of $m-n$, and on this basis obtained a general formula for $\max(S_n/E_n)$.

The first published proof of Clenshaw's conjecture was given by Lam and Elliott [8] in 1972. Using the same method and assumption as Clenshaw, they were able to generalize his results to any value of $m-n$, although they failed to consider the important question of whether the error of approximation must always be representable in the form they assumed for it. This omission was remedied in their recent second paper [7], in which not only is this question considered, but the norm of error is shown to be given by an eigenvalue of a certain matrix. That this should be so is not at all surprising however, for as is clear from the author's papers [10] and [11], any problem of uniform approximation of polynomials or rational functions by polynomials or rational functions is likely to lead to an eigenvalue problem.

The present paper uses a simplified form of the "u-method" developed in [10, 11], to deal with the problem treated by Lam and Elliott. Our treatment differs significantly from theirs; we use standard results from approximation theory rather than matrix theorems. Not only does this lead to some simplification, but it also provides a proof that the desired solution exists *unconditionally*. A proof of Clenshaw's conjecture is also given.

2. Preliminary Discussion

We denote the given polynomial of degree m by $f(x)$, and for convenience write $m=n+r+1$. We take the given interval as $[-1,1]$. Let $f(x)$ have the expansion (with $a_r \neq 0$)

$$f(x)=a_rT_m(x)+a_{r-1}T_{m-1}(x)+\cdots+a_0T_{n+1}(x)+\text{lower order terms.} \tag{1}$$

Then the error norm S_n of the truncated expansion is

$$S_n=\| a_rT_m+\cdots+a_0T_{n+1}\| \tag{2}$$

where $\|\cdot\|$ denotes maximum modulus in $[-1,1]$. The error norm E_n of the best nth degree polynomial approximation $\hat{P}_n$ to f is

$$E_n=\inf_{P\in\mathbb{P}_n}\| P-f\|=\|\hat{P}_n-f\| \tag{3}$$

where $\mathbb{P}_n$ denotes the set of all real polynomials of degree $\leqslant n$. We note that by the Alternation theorem, $\hat{P}_n-f=\pm E_n$ alternately at $n+2$ or more points on $[-1,1]$.

Instead of finding $\hat{P}_n$ we shall obtain an infinite set of rational functions Q/D, where $Q \in \mathbb{P}_{n+r}$, $D \in \mathbb{P}_r$, with error function

$$R = (Q/D) - f = M/D, \qquad M = Q - Df \tag{4}$$

such that $R = \pm \| R \|$ alternately at $n + 2$ or more points on $[-1, 1]$. A unique member of this set is of course the "best" or *optimal* rational approximation, i.e., that which minimizes $\| R \|$ for all possible choices of $Q \in \mathbb{P}_{n+r}$, $D \in \mathbb{P}_r$. As is well known, this R exhibits alternation not merely at $n + 2$ points but in general at $n + 2r + 2$ points. We shall show that another likewise unique member of the set has instead the special property that $\| \mathrm{Fr}(Q/D) \| \to 0$ as $n \to \infty$, where Fr denotes "fractional part." Its integral part is then the desired polynomial approximation to f. We shall call this member of the set the "asymptotic Q/D." In order to prove its existence we shall demonstrate a close "dual" relationship between the desired function Q/D and the optimal rational approximation to a certain polynomial g of degree m related in a special way to f. Since the algebraic solution for the asymptotic Q/D is exactly the same as for the optimal Q/D, we shall start by considering the problem of finding the optimal approximation Q/D to f.

If for this optimum, expressed in its lowest terms, the actual degrees of D and Q are, respectively, $s = r - d$ and $n + s' = n + r - d'$, where d, $d' \geqslant 0$, the problem has "deficiency" $\delta = \min(d, d')$, and by the Alternation theorem for rational approximation (see for example Rivlin [9, Theorem 5.2], $R = \pm \| R \|$ alternately at $\kappa = n + 2r + 2 - \delta$ or more points on $[-1, 1]$. Let $E = \| R \|$. Then $R^2 - E^2$ has at least κ distinct zeros in $[-1, 1]$, of which $\tau \leqslant 2$ are at the end points ± 1 and $\kappa - \tau$ are internal and of order at least 2. Thus $M^2 - E^2 D^2$ has at least $2(\kappa - \tau) + \tau = 2\kappa - \tau$ zeros in $[-1, 1]$, counting multiplicities. But its degree is $2(m + s) \leqslant 2(\kappa - 1)$, since $s \leqslant r - \delta$. It follows that $\tau = 2$, i.e., $R = \pm E$ at both end points, and that $\delta = d$, i.e., $s' \leqslant s$, so that Q has degree at most $n + s$. Further, $M^2 - E^2 D^2$ has precisely $\kappa - 2$ internal zeros of order 2, and no external zeros. We may therefore write, noting that $M^2 - E^2 D^2 \leqslant 0$ in $[-1, 1]$,

$$M^2 - E^2 D^2 = (x^2 - 1)\, W^2, \tag{5}$$

where W is real, of degree $n + r + s$, and has all its roots in $(-1, 1)$.

It is clear that if M, D, W is any triplet of real polynomials satisfying (5) for some value of E, then $\| M/D \| = E$. We shall see that if we start with any suitable D (viz, real D as in (10) below) we can obtain many such triplets. In general the corresponding $Q = M + Df$ will have degree $m + s$, but as we shall see, if we impose the condition found above that Q shall have nominal degree $n + s$ instead of $m + s$, then we can obtain both the optimal Q/D and the asymptotic Q/D which we seek. In Section 3, we obtain general solutions of (5) and consider the implications of the desired asymptotic property, in Section 4 we involve the given f explicitly by imposing the degree condition on Q, and in Section 6 we use the existence of the optimal Q/D to establish the existence of the Q/D sought.

The method used is that already described in the author's earlier papers [10, 11]. However, a key step in the process, namely, the factorization of (18) below, is treated much more simply here than in those papers, where the treatment was based on the rather complicated surd factorization theorem in [10]. For the sake of completeness the method is described in full, reference to [10] being made only at one point in Section 4.

Remark. The method to be described requires (5) or an equivalent equation as a starting point. Unfortunately best-approximation problems involving polynomials and rational functions do not always lead to equations of this form. For example, in the case $r = 1$ (i.e., the case solved by Zolotarev) the optimal error function satisfies an equation either of the form

$$R^2 - E^2 = (x + 1)(x - \beta)\, W^2$$

(so that only one of the end points is a "norm-point") which is reducible to the form (5) by a simple linear transformation in x; or of the form

$$R^2 - E^2 = (x^2 - 1)(x - \alpha)(x - \beta)\, W^2$$

which requires elliptic functions for its solution, and cannot be dealt with by the present method.

3. General solutions of (5)

We rewrite (5) as

$$M^2 - (x^2 - 1)\, W^2 = E^2D^2 \tag{6}$$

and make the left-hand side factorizable by means of the substitution

$$x = \tfrac{1}{2}(u + u^{-1}), \tag{7}$$

giving

$$x^2 - 1 = \tfrac{1}{4}(u - u^{-1})^2. \tag{8}$$

We note that

$$T_k(x) = \tfrac{1}{2}(u^k + u^{-k}), \qquad k = 0, 1, 2,.... \tag{9}$$

Now to allow for possibly degenerate solutions we suppose D has degree $s = r - d$, $d \geqslant 0$, say

$$D = \prod_1^s (x - x_j), \qquad x_j \notin [-1, 1]. \tag{10}$$

Each x_j can be expressed as

$$x_j = \tfrac{1}{2}(u_j + u_j^{-1}), \tag{11}$$

where there are two possible values of u_j, reciprocals of each other.

Combining (7) and (11) gives

$$x - x_j = (-1/2u_j)(u - u_j)(u^{-1} - u_j), \tag{12}$$

so that if we write

$$\phi(u) = \prod_1^s (u - u_j) = \phi_s u^s + \cdots + \phi_1 u + \phi_0 \tag{13}$$

we have

$$D = (1/2^s \phi_0)\, \phi(u)\, \phi(u^{-1}). \tag{14}$$

We now define

$$p(u) = M(x) + \tfrac{1}{2}(u - u^{-1})\, W(x), \tag{15}$$

with the sign of W chosen so that $p(u)$ is of order $O(u^{m+s})$ for large u (i.e., there is no cancellation of leading terms in (15)). Then

$$p(u^{-1}) = M(x) - \tfrac{1}{2}(u - u^{-1})\, W(x). \tag{16}$$

and we have

$$M = \tfrac{1}{2}(p(u) + p(u^{-1})), \tag{17}$$

while by (6)

$$p(u)\, p(u^{-1}) = (E/2^s \phi_0)^2\, \phi^2(u)\, \phi^2(u^{-1}). \tag{18}$$

Now by (15) and (16) $p(u)$ and $p(u^{-1})$ have no poles except possibly at $u = 0$, so by (18) they can have no zeros except at $u = 0$ and at zeros of $\phi(u)$, $\phi(u^{-1})$. There are then only two distinct possibilities arising from (18):

(a) $p(u) = (-\lambda/2^s \phi_0)\, \phi^2(u)\, u^\tau$, ($\lambda = \pm E$, τ an integer),

or

(b) $p(u) = (-\lambda/2^s \phi_0)\, \phi_1(u)\, \phi_1(u^{-1})\, \phi_2^2(u)\, u^\tau$, where $\phi_1(u)\, \phi_2(u) = \phi(u)$.

Case (b) leads to a solution in which M, W, and D automatically have common factors, and we disregard this as it is not needed. An apparent modification of (a) in which some or all of the factors of $\phi(u)$ are replaced by corresponding factors of $\phi(u^{-1})$ is easily seen to lead to the same solution as (a), bearing in mind that

$$u^{-1} - u_j = -u_j u^{-1}(u - u_j^{-1}),$$

where u_j^{-1} is an alternative choice for u_j, for a given x_j. Thus we shall take (a) as our expression for $\phi(u)$. Since $p(u) = O(u^{m+s})$, it follows at once that

$$\tau = m - s = n + 1 + d.$$

Thus

$$p(u) = (-\lambda/2^s\phi_0)\, u^{n+1+d}\phi^2(u) \qquad (\lambda = \pm E, E = |\lambda|), \tag{19}$$

and

$$M(x) = (-\lambda/2^{s+1}\phi_0)(u^{n+1+d}\phi^2(u) + u^{-n-1-d}\phi^2(u^{-1})). \tag{20}$$

Now if the x_j are all distinct (and otherwise a continuity argument may be used),

$$\mathrm{Fr}\left(\frac{Q}{D}\right) = \mathrm{Fr}\left(\frac{M}{D}\right) = \sum \frac{M(x_j)}{(x - x_j)\, D'(x_j)}$$

where

$$M(x_j) = \frac{-\lambda}{2^{s+1}\phi_0}\, u_j^{-n-1-d}\phi^2(u_j^{-1}).$$

As will be seen in Section 4, the u_j and λ depend only on the $r + 1$ prescribed leading coefficients in the expansion (1) of f[1], and not at all on n. Hence

$$\| \mathrm{Fr}(Q/D) \| \to 0 \text{ as } n \to \infty \text{ if and only if all } | u_j | > 1.^{2} \tag{21}$$

Now any solution Q/D, after reduction if necessary to lowest terms, corresponds to M and D without common factors. This means by (14) and (20) that $\phi(u)$ and $\phi(u^{-1})$ then have no common factor, in other words $| u_j | \neq 1$

[1] The remaining coefficients are unimportant, for they contribute merely an additive polynomial of degree n to the solution.

[2] Obviously for any given D satisfying (10) the u_j can be chosen to satisfy the condition $| u_j | > 1$. However, we still have to impose the degree condition on Q, which in fact takes the form (26), and so determines $\phi(u)$ rather than $D(x)$. Thus, the choice of u_j is not at our disposal, and it will be our task in Section 5 to show that there is a solution $\phi(u)$ of (26) which satisfies the condition in (21).

for all j. (Note that this also implies that our solutions, when in lowest terms, satisfy the condition on the x_j in (10). This follows alternatively directly from (6).) Thus if we denote by β the number of zeros u_j of $\phi(u)$ inside the unit circle, the asymptotic property sought will be achieved if $\beta = 0$.

We now derive a simple general relation between β and the number α of alternation points on $[-1, 1]$ (i.e., points at which $R = \pm E$ alternately.) For this, we note that the transformation (7) maps the semicircle $u = e^{i\theta}$, $0 \leqslant \theta \leqslant \pi$ onto the interval $1 \geqslant x \geqslant -1$, where $x = \cos\theta$. On moving round the semicircle, we have by (19)

$$\Delta \arg p = (n + 1 + d)\pi + 2\beta\pi.$$

On the other hand, we have on the semicircle

$$p(u) = M + iW \sin\theta,$$

where M and W are real, so that

$$\Delta \arg p = (\alpha - 1)\pi.$$

It follows that

$$\alpha = n + 2 + d + 2\beta. \tag{22}$$

Thus for any solution of (5) in which D has degree $s = r - d$ the number of alternation points must be at least $n + 2 + d$, and for the asymptotic solution we seek (if it exists) for which $\beta = 0$ the number is precisely $n + 2 + d$, i.e., in the case $d = 0$ the same as for the optimal polynomial $\hat{P}_n$. For the optimal rational function on the other hand the number is at least $\kappa = n + 2 + 2r - d$, so that for this solution β must be equal to s, its maximum possible value.

We have thus exhibited a kind of inverse relationship between the optimal and the asymptotic Q/D. We shall see in Section 6 that there is a further relationship between these two through which we can prove the existence of the asymptotic Q/D. Now assuming this for a moment, suppose that the alternation points in $[-1, 1]$ are $y_1, y_2, \ldots$ in ascending order. Then if $P = \operatorname{Int}(Q/D)$,

$$P(y_k) - f(y_k) = \epsilon(-1)^k E - \operatorname{Fr}(Q/D)(y_k), \qquad k = 1, \ldots, n + 2 + d,$$

where $\epsilon = \pm 1$. Thus if $\| \operatorname{Fr}(Q/D) \| = \nu < E$, $P - f$ alternates in sign at the y_k, and, using de la Vallée Poussin's theorem (e.g., Cheney [5, p. 77]) accordingly,

$$E - \nu \leqslant E_{n+d} \leqslant E_n \leqslant \| P - f \| \leqslant E + \nu. \tag{23}$$

Now as we have seen, $\nu \to 0$ as $n \to \infty$, and moreover, because of the form of f in (1), if a_p is the first of $a_0, a_1, \ldots,$ which is nonzero, $E_n \geqslant E_{n+p} \geqslant | a_p |$, if $n \geqslant r/2$ (see [5, p. 137 Theorem 5]). It follows that, for fixed r and fixed $a_0, \ldots, a_r$,

$$E/E_n \to 1 \qquad \text{as } n \to \infty. \tag{24}$$

Remark. When $\phi(u)$ is known, the internal norm-points of error, where $R = +E$ or $-E$, may be found jointly as the roots of $W(x) = 2(p(u) - p(u^{-1}))/(u - u^{-1})$. They may however be found separately as roots of two polynomials each of about half the degree of W, for by (4), (14) and (20)

$$R \mp \lambda = \frac{-\lambda}{2^{s-1}\phi_0} \frac{\omega_{\pm}{}^2}{D}, \tag{25}$$

where

$$\omega_{\pm} = \tfrac{1}{2}(u^{(1/2)(n+d+1)}\phi(u) \pm u^{-(1/2)(n+d+1)}\phi(u^{-1})).$$

If $n + d$ is odd, say $2h - 1$, the identities

$$u^k + u^{-k} = 2T_k(x), \qquad u^k - u^{-k} = 2(x^2 - 1)^{1/2}\, U_{k-1}(x)$$

give

$$\omega_{+} = \phi_0 T_h + \cdots + \phi_s T_{h+s},$$
$$\omega_{-} = (x^2 - 1)^{1/2}\, (\phi_0 U_{h-1} + \cdots + \phi_s U_{h+s-1}),$$

while if $n + d$ is even, say $2h$, the identities

$$u^{k+(1/2)} \pm u^{-k-(1/2)} = [2(x \pm 1)]^{1/2}\, (U_k(x) \mp U_{k-1}(x))$$

give

$$\omega_{\pm} = [\tfrac{1}{2}(x \pm 1)]^{1/2}\, (\mp \phi_0 U_{h-1} + (\phi_0 \mp \phi_1)\, U_h + \cdots$$
$$+ (\phi_{s-1} \mp \phi_s)\, U_{h+s-1} + \phi_s U_{h+s}).$$

The internal norm-points are roots of $\omega_{\pm}$ other than 1 or -1. They may of course also be found directly from $\omega_{\pm}$ above as functions of u.

4. Solutions for Given $f(x)$

We have so far considered solutions of (5) with arbitrary D of degree s (and nonzero on $[-1, 1]$), but without reference to $f(x)$. We must now impose the condition that $Q \equiv M + Df$ has degree $n + s$ (or less) instead of

$m + s = n + s + r + 1$. Now for large u we have, on dividing by $\phi(u)$,

$$\begin{aligned}\frac{M + Df}{\phi(u)} = {} & \frac{-\lambda}{2^{s+1}\phi_0}\left(u^{n+d+1}\phi(u) + O(u^{-m})\right) \\ & + \frac{1}{2^{s+1}\phi_0}(\phi_0 + \phi_1 u^{-1} + \cdots \phi_s u^{-s}) \\ & \times (a_r u^m + a_{r-1}u^{m-1} + \cdots + a_0 u^{n+1} + \cdots)\end{aligned}$$

while $Q/\phi(u) = O(u^n)$. Equating coefficients of $u^{n+1}, u^{n+2}, \ldots, u^m$ on both sides of the equation

$$(M + Df)/\phi(u) = Q/\phi(u)$$

gives a set of equations which may be written

$$\begin{bmatrix} a_0 & a_1 & a_2 & \cdot & \cdot & \cdot & a_{r-1} & a_r \\ a_1 & a_2 & & & & \cdot & a_r & \\ a_2 & & & & \cdot & \cdot & & \\ \cdot & & & \cdot & \cdot & & & \\ \cdot & & \cdot & \cdot & & & & \\ \cdot & \cdot & \cdot & & & & & \\ a_{r-1} & a_r & & & & 0 & & \\ a_r & & & & & & & \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \cdot \\ \cdot \\ \phi_s \\ 0 \\ \cdot \\ 0 \end{bmatrix} = \lambda \begin{bmatrix} 0 \\ \cdot \\ 0 \\ \phi_0 \\ \phi_1 \\ \cdot \\ \cdot \\ \phi_s \end{bmatrix} \tag{26}$$

or briefly

$$A\phi_{(d)} = \lambda S^d \phi_{(d)} \tag{27}$$

where A is the $(r + 1) \times (r + 1)$ triangular Hankel matrix shown, $\phi_{(d)}$ is an $(r + 1)$-element vector consisting of the $s + 1$ coefficients of $\phi(u)$ (forming a vector ϕ, say) supplemented by d zero elements, and S is a shifting matrix defined by

$$\begin{aligned}(S)_{ij} &= 1 \qquad \text{if } i = j + 1 = 2, \ldots, r + 1, \\ &= 0 \qquad \text{otherwise.}\end{aligned}$$

Now it is clear by inspection of (26) that if A_h is the matrix (with leading element a_h and of similar form to A) obtained from A by deleting the first h rows and the last h columns (so that $A_0 = A$), then (26) implies that

$$A_{d-k}\phi_{(k)} = \lambda S^k \phi_{(k)}, \qquad k = 0, 1, \ldots, d. \tag{28}$$

Thus in particular λ is an eigenvalue of A_d and ϕ an eigenvector. It is now obvious that with such λ and ϕ, if $d > 0$ the first d equations in (26) will not in general be satisfied. Thus in general we must have $d = 0$, and λ an eigenvalue of A with eigenvector ϕ.

In exceptional (degenerate) cases however (27) may have a solution for some $d > 0$. It then follows (as was shown in [10]) that the equations

$$A\mathbf{y}_{(k)} = \lambda S^k \mathbf{y}_{(k)}, \tag{29}$$

$$A\mathbf{z}_{(k)} = -\lambda S^k \mathbf{z}_{(k)}\,, \tag{30}$$

both have solutions for $k = d - 1, d - 2, \ldots, 1, 0$. (These solutions correspond to multiplying $\phi(u)$ by one or more of $1 + u$, $1 - u$, or factors of the form $1 + 2cu + u^2$ (with arbitrary c), and hence D, M, and Q by one or more common factors $x + 1, x - 1$ or $(x + c)^2$.) In particular, taking $k = 0$, it follows that if (27) has a solution for some $d > 0$, then although $\phi_{(d)}$ is not an eigenvector of A, both λ and $-\lambda$ must be eigenvalues of A. (In fact, as was shown in [10], λ is an eigenvalue of order at least $[\frac{1}{2}(d + 2)]$ and $-\lambda$ of order at least $[\frac{1}{2}(d + 1)]$.)

For eigenvalues λ of A we shall denote by $d(\lambda)$ the largest value of d for which (27) has a solution; for non-eigenvalues λ it is convenient to let $d(\lambda) = -1$. Now (27) has a solution if and only if the first $r + 1 - d = s + 1$ columns of $A - \lambda S^d$ are linearly dependent, i.e., the matrix

$$\begin{bmatrix}
a_0 & a_1 & \cdot & \cdot & \cdot & \cdot & a_s \\
\cdot & & & & & & \cdot \\
\cdot & & & & & & \cdot \\
a_{d-1} & & & & & & a_{r-1} \\
a_d - \lambda & \cdot & \cdot & \cdot & \cdot & \cdot & a_r \\
a_{d+1} & a_{d+2} - \lambda & & & & \cdot & \\
\cdot & & \cdot & & \cdot & & \\
\cdot & & & a_r - \lambda & & & 0 \\
\cdot & & \cdot & & -\lambda & & \\
\cdot & \cdot & & & & \cdot & \\
a_r & & & 0 & & & -\lambda
\end{bmatrix} \tag{31}$$

has rank at most s. Thus if λ is an eigenvalue of A, $d(\lambda)$ is the maximum value of d for which this is true.

From the results above it is clear that for any λ,

$$| d(\lambda) - d(-\lambda)| \leqslant 1. \tag{32}$$

We note next that just as (27) implies (28), so (29) and (30) imply

$$A_{k-h}\,\mathbf{y}_{(h)} = \lambda S^h\,\mathbf{y}_{(h)}, \tag{33}$$

$$0 \leqslant h \leqslant k \leqslant d - 1;$$

$$A_{k-h}\,\mathbf{z}_{(h)} = -\lambda S^h\,\mathbf{z}_{(h)}, \tag{34}$$

whence λ and $-\lambda$ are eigenvalues of all A_k, $0 \leqslant k \leqslant d-1$, if (27) holds, while as we have seen λ is also an eigenvalue of A_d. We note also that since by assumption $a_r \neq 0$, the A_k are nonsingular for all k, and $\lambda \neq 0$.

It is clear that if for some $p \geqslant 0$, λ is an eigenvalue of $A, A_1, \ldots, A_p$ but not of A_{p+1}, then $0 \leqslant d(\lambda) \leqslant p$. It might be surmised that in fact $d(\lambda) = p$, but in general this will not be true, as may be easily seen by noting that the only term containing a_0 in the expansion of $\det(A - \lambda I)$ is $(a_0 - \lambda)\det(A_2 - \lambda I)$. Thus if λ is an eigenvalue of A_2 and A, it remains an eigenvalue of A even if a_0 is varied, whereas (27) cannot continue to hold, i.e., (31) to have less than full rank when $d = p$, as a_0 is varied, unless λ is an eigenvalue of A_{p+2}, which cannot be the case since by Lemma 1 below it would imply that λ is an eigenvalue of A_{p+1}.

We proceed to prove this Lemma, and make use of it in proving two further lemmas relating to the cases $d(\lambda) = 0$ and $d(\lambda) = 1$. It is convenient to use the notation

$$D_p(\lambda) = \det(A_p - \lambda I), \qquad D(\lambda) = \det(A - \lambda I) = D_0(\lambda).$$

LEMMA 1. $D_p(\lambda) = D_{p+2}(\lambda) = 0 \Rightarrow D_{p+1}(\lambda) = 0$.

Proof. Let λ be such that $D_p(\lambda) = D_{p+2}(\lambda) = 0$. Then corresponding to the eigenvalue λ of A_p there is an eigenvector $(x_p, x_{p+1}, \ldots, x_r)'$ with $x_p = x_r = 0$, in other words the columns of $A_p - \lambda I$ other than the first and last are linearly dependent. For consider the cofactors of top-row elements in $\det(A_p - \lambda I)$. Those of $a_p - \lambda$ and a_r, namely, $-\lambda D_{p+2}(\lambda)$ and $\pm a_r D_{p+2}(\lambda)$, are both zero. If any of the remainder are nonzero, we can take the set of cofactors as our eigenvector elements, since $D_p(\lambda) = 0$. If all are zero, the rows of $A_p - \lambda I$ after the first are linearly dependent, with multipliers $m_{p+1}, \ldots, m_r$, say, and by symmetry the same applies to the columns. But since $\lambda \neq 0$ it is obvious by inspection of the last row that $m_r = 0$, and we may take $(0, m_{p+1}, \ldots, m_{r-1}, 0)'$ as our eigenvector. (Alternatively, since $a_r x_p = \lambda x_r$ for an eigenvector, $x_p = 0$ if and only if $x_r = 0$.) In either case, $x_p = x_r = 0$.

Now it is easy to verify that

$$A_{p+1}(0, x_{p+1}, \ldots, x_{r-1})' = \lambda(x_{p+1}, \ldots, x_{r-1}, 0)',$$

$$A_{p+1}(x_{p+1}, \ldots, x_{r-1}, 0)' = \lambda(0, x_{p+1}, \ldots, x_{r-1})'.$$

It follows that λ is an eigenvalue of A_{p+1}, with eigenvector

$$(x_{p+1}, x_{p+1} + x_{p+2}, \ldots, x_{r-2} + x_{r-1}, x_{r-1})'. \tag{35}$$

Lemma 2. *If λ is an eigenvalue of A_p, with eigenvector $(x_p, \dots, x_r)'$, and $D_{p+1}(\lambda) \neq 0$, then $x_p \neq 0, x_r \neq 0$.*

Proof. By Lemma 1, $D_{p+2}(\lambda) = 0$. Thus the cofactor of $a_p - \lambda$ in $A_p - \lambda I$ is nonzero, while $\det(A_p - \lambda I) = 0$. It follows that the space of solutions of $A_p\mathbf{x} = \lambda\mathbf{x}$ has dimension 1, and any solution has elements proportional to the cofactors of top-row elements of A_p. In particular $x_p \neq 0$, which implies $x_r \neq 0$.

We may note that in the case $p = 0$, i.e., when $D(\lambda) = 0$, $D_1(\lambda) \neq 0$, we have $d(\lambda) = 0$.

Lemma 3. *Let $D_p(\lambda) = D_{p+1}(\lambda) = 0$, $D_{p+2}(\lambda) \neq 0$. Then if $\mathbf{x}$ is an eigenvector of A_{p+1} corresponding to λ,*

$$A_p\mathbf{x}_{(1)} = \lambda S\mathbf{x}_{(1)} . \tag{36}$$

Proof. For simplicity we shall prove this for the case $p = 0$: the result immediately generalizes for any $p > 0$. We assume then that $D(\lambda) = D_1(\lambda) = 0$, $D_2(\lambda) \neq 0$.

By Lemma 1, $D_3(\lambda) \neq 0$. Since the only term in $D_1(\lambda)$ containing a_1 is $(a_1 - \lambda)\, D_3(\lambda)$, the equation $D_1(\lambda) = 0$ is equivalent to

$$a_1 = a_1(a_2, \dots, a_r, \lambda), \tag{37}$$

where $a_1(\ )$ is a certain rational function in the variables.

Similarly, with $D_2(\lambda) \neq 0$, $D(\lambda) = 0$ is equivalent to

$$a_0 = a_0(a_1, a_2, \dots, a_r, \lambda)$$

which when combined with (37) gives

$$a_0 = \tilde{a}_0(a_2, \dots, a_r, \lambda). \tag{38}$$

Now let $A_1\mathbf{x} = \lambda\mathbf{x}$, $\mathbf{x} = (x_1, \dots, x_r)'$. Here $x_1 \neq 0$, by Lemma 2. Then (36) will hold provided the additional condition

$$a_0x_1 + a_1x_2 + \cdots + a_{r-1}x_r = 0 \tag{39}$$

is satisfied. Since $x_1, \dots, x_r$ are all expressible as polynomials in $a_2, \dots, a_r, \lambda$, with $x_1 \neq 0$, (39) when combined with (37) is equivalent to

$$a_0 = \bar{a}_0(a_2, \dots, a_r, \lambda). \tag{40}$$

Further, if (36) holds then λ is an eigenvalue of A. We have therefore the following sequence of implications:

$$\begin{aligned} &D_2(\lambda) \neq 0,\ D_3(\lambda) \neq 0,\ (37) \text{ and } (40) \\ &\quad \Rightarrow D_1(\lambda) = 0,\ D_2(\lambda) \neq 0,\ (37) \text{ and } (40) \\ &\quad \Rightarrow A_1\mathbf{x} = \lambda\mathbf{x} \text{ (for some } \mathbf{x}),\ D_2(\lambda) \neq 0,\ (37) \text{ and } (39) \\ &\quad \Rightarrow (36),\ D_2(\lambda) \neq 0, \text{ and } (37) \\ &\quad \Rightarrow D(\lambda) = 0,\ D_2(\lambda) \neq 0, \text{ and } (37) \\ &\quad \Rightarrow (38). \end{aligned}$$

Thus for almost arbitrary λ, $a_2, \ldots, a_r$ (restricted only by the conditions $D_2(\lambda) \neq 0$, $D_3(\lambda) \neq 0$), (40) implies (38). We can therefore conclude that the functions $\tilde{a}_0$ and $\bar{a}_0$ are identical and we may now write

$$\begin{aligned} &D(\lambda) = 0,\ A_1\mathbf{x} = \lambda\mathbf{x},\ D_2(\lambda) \neq 0 \\ &\quad \Rightarrow D_2(\lambda) \neq 0,\ D_3(\lambda) \neq 0,\ (37) \text{ and } (38) \\ &\quad \Rightarrow D_2(\lambda) \neq 0,\ D_3(\lambda) \neq 0,\ (37) \text{ and } (40) \\ &\quad \Rightarrow (36), \end{aligned}$$

which proves the theorem.

We note that in the case $p = 0$, $d(\lambda) = 1$.

As a corollary of Lemma 3 we have:

LEMMA 4. *If* $p \geqslant 1$ *and* $D_p(\lambda) = D_{p+1}(\lambda) = 0$, $D_{p+2}(\lambda) \neq 0$, *then* $D_{p-1}(\lambda) = 0$.

Proof. By Lemma 3,

$$A_p(x_{p+1}, \ldots, x_r, 0)' = \lambda(0, x_{p+1}, \ldots, x_r)'.$$

It immediately follows that

$$A_{p-1}(0, x_{p+1}, \ldots, x_r, 0)' = \lambda(0, x_{p+1}, \ldots, x_r, 0)', \tag{41}$$

whence λ is an eigenvalue of A_{p-1}.

It is obvious that in the case $p = 1$, $d(\lambda) \leqslant 2$. In general however we will have $d(\lambda) = 0$ in this case, since a_0 is arbitrary.

5. OPTIMAL RATIONAL APPROXIMATION

We consider now the optimal Q/D for the given $f(x)$, with $Q \in \mathbb{P}_{n+r}$, $D \in \mathbb{P}_r$. We know from the existence theorem for rational approximation that there exists a unique optimum, say $\hat{Q}/\hat{D}$ in lowest terms, and we have seen in Section 2 that if $\hat{D}$ has actual degree $s = r - d$ the problem has "deficiency" $\delta = d$, and $\hat{Q}$ has actual degree $n + s' \leqslant n + s$. Further, the optimum must satisfy (5) and hence (27). Now any solution of (27) with any value of d yields Q/D with error norm $E = | \lambda |$, where λ is an eigenvalue of A; while on the other hand any eigenvalue and its eigenvector satisfy an equation of form (27) with $d = 0$. Since the optimum has minimum error norm for all Q/D considered, it follows that its error norm is

$$\hat{E} = \min | \lambda | \tag{42}$$

taken over all eigenvalues λ of A.

If the deficiency is δ, (27) holds with $d = \delta$ and $\lambda = \hat{E}$ or $-\hat{E}$ (but not both since the optimum is unique), and if $\delta > 0$, (29) and (30) hold with $k = 0, 1,..., \delta - 1$. Moreover (27) cannot hold with $\lambda = \hat{E}$ or $-\hat{E}$ for $d > \delta$, since then (29) and (30) would both hold with $k = \delta$, and the optimum would not be unique. Since A, being symmetric, has a full set of distinct eigenvectors even if some of its eigenvalues are multiple we have proved the following theorem:

THEOREM 1. *The unique rational approximation Q/D to $f(x)$ on* $[-1, 1]$, *with $Q \in \mathbb{P}_{n+r}$, $D \in \mathbb{P}_r$ and $f(x)$ given as in* (1), *has error norm $\hat{E}$ = minimum eigenvalue modulus of the matrix A in* (26), *and actual degree of D equal to* $r - \delta$, where δ *is the deficiency.*

If the eigenvalue of minimum modulus is unique, $\delta = 0$ (*and conversely.*) *Otherwise both $\hat{E}$ and $-\hat{E}$ are eigenvalues of A and*

$$\delta = \max(d(\hat{E}), d(-\hat{E})), \tag{43}$$

i.e., δ is the largest d for which (27) *has a solution with $\lambda = \hat{E}$ or $-\hat{E}$, there being only one such solution for $d = \delta$* (i.e., $d(\hat{E}) \neq d(-\hat{E})$.)

An upper bound on δ can be found from the orders of the eigenvalues $\hat{E}$ and $-\hat{E}$ of A. If these are p and q, then as already noted $[\frac{1}{2}(\delta + 2)] \leqslant p$ and $[\frac{1}{2}(\delta + 1)] \leqslant q$, or vice versa. It follows that

$$\delta \leqslant 2 \min(p, q) \text{ if } p \neq q, \qquad \delta \leqslant 2p - 1 \text{ if } p = q. \tag{44}$$

It is important to note that since $D(x)$ and therefore $\phi(u)$ has actual degree $s = r - \delta$, our solution $\phi(u)$ has $\phi_s \neq 0$, from which it follows by (27) that $\phi_0 \neq 0$, and that ϕ_s can be normalised to 1 as in (13). Further, the matrix (31), with $d = \delta$, has rank precisely s, as the solution is unique.

We note also that since δ and thus s is determined through A from $f(x)$ by (43), the Q/D we obtain must be in lowest terms.

Remark. If $\delta > 0$, each eigenvector of A corresponding to the simple or multiple eigenvalues $\hat{E}$ and $-\hat{E}$ yields a solution Q/D with common factors in Q and D (viz, $x + 1$, $x - 1$, or $(x + c)^2$ for some c). On cancelling these, Q/D reduces necessarily to the unique optimal Q/D. Thus in fact the optimum can always be found without regard to the deficiency, i.e., by proceeding as if $\delta = 0$, and using either $\hat{E}$ or $-\hat{E}$. This will however lead to heavier calculations.

6. Existence of Asymptotic Solution

To complete our analysis of the problem we have to show that among the solutions of (27) there is at least one giving $\phi(u)$ with all its roots outside the unit circle. To do this we exploit still further the dual relationship already noted between the asymptotic solution we seek, for which $\beta = 0$, and the optimal solution, for which $\beta = s$. The key to our proof is the observation, easily proved by induction, that A has an inverse A^{-1} which is of similar form when reflected in the secondary diagonal, i.e.,

$$A^{-1} = \begin{bmatrix} & 0 & & & & b_r \\ & & & & \cdot & \cdot \\ & & & \cdot & & \cdot \\ & & \cdot & & & b_2 \\ & \cdot & & & b_2 & b_1 \\ b_r & \cdot & \cdot & b_2 & b_1 & b_0 \end{bmatrix}. \qquad (45)$$

Thus if we denote by P the unit matrix with its columns (or rows) reversed, then $B = PA^{-1}P$ has the same form as A, and corresponds to a given polynomial $g(x)$ "dual" to $f(x)$:

$$g(x) = b_r T_m(x) + b_{r-1} T_{r-1}(x) + \cdots + b_0 T_{n+1}(x) + \cdots. \qquad (46)$$

We note that the eigenvalues of B are the reciprocals of those of A.

Now the unique optimal Q/D for $g(x)$ is governed by Theorem 1, with A replaced by B, and (27) replaced by

$$B\psi_{(d)} = \mu S^d \psi_{(d)} \qquad (47)$$

say, where we are using the notation ψ instead of ϕ, and μ instead of λ. Then if (47) holds for some μ and $\psi_{(d)}$, and we write

$$\phi_{(d)} = (\phi_0, \ldots, \phi_s, 0, \ldots, 0)' = (\psi_s, \ldots, \psi_0, 0, \ldots, 0)', \qquad (48)$$

we have

$$P\boldsymbol{\phi}_{(d)} = S^d \boldsymbol{\psi}_{(d)}, \qquad P\boldsymbol{\psi}_{(d)} = S^d \boldsymbol{\phi}_{(d)} \tag{49}$$

and

$$A\boldsymbol{\phi}_{(d)} = \lambda S^d \boldsymbol{\phi}_{(d)}, \tag{50}$$

where $\lambda = 1/\mu$. Conversely (50) implies (47). Thus (47) and (50) are equivalent dual relationships, linked by (48), and to any eigenvalue μ of B for which (47) holds corresponds an eigenvalue $\lambda = 1/\mu$ of A for which (50) holds.

The optimal Q/D for g has deficiency δ equal to the largest d for which (47) has a solution $\boldsymbol{\psi}_{(d)}$ when μ is an eigenvalue of B of minimum modulus. By the duality it is clear that δ is also equal to the largest d for which (50) has a solution $\boldsymbol{\phi}_{(d)}$ when λ is an eigenvalue of A of *maximum* modulus.

Now the optimal $\boldsymbol{\psi}_{(\delta)}$ yields a polynomial $\psi(u) = \psi_0 + \psi_1 u + \cdots + \phi_s u^s$, with $s = r - \delta$, having all its roots inside the unit circle ($\beta = s$). The dual vector $\boldsymbol{\phi}_{(\delta)}$ yields $\phi(u) = \psi_s + \psi_{s-1} u + \cdots + \psi_0 u^s = u^s \psi(1/u)$, which has all its roots outside the unit circle ($\beta = 0$). We have thus established the existence of the asymptotic solution. Its uniqueness follows from that of the optimal Q/D, which is *characterized* by the condition $\beta = s$.

Further, as already noted in Section 5, any solution of (50) with any value of d yields Q/D with error norm $E = \| Q/D - f \| = | \lambda |$. Thus the "asymptotic" Q/D we have found has error norm $\tilde{E}$ equal to the largest eigenvalue modulus of A:

$$\tilde{E} = \max | \lambda | . \tag{51}$$

We can now state the analog of Theorem 1:

THEOREM 2. *There is a unique "asymptotic" rational approximation Q/D to $f(x)$ on $[-1, 1]$, with $Q \in \mathbb{P}_{n+r}$, $D \in \mathbb{P}_r$ and $f(x)$ given as in* (1). *It has error norm $\tilde{E}$ = maximum eigenvalue modulus of the matrix A in* (26), *and actual degree of D equal to $r - \delta$, where δ is the deficiency.*

If the eigenvalue of maximum modulus is unique, $\delta = 0$ (and conversely). Otherwise both $\tilde{E}$ and $-\tilde{E}$ are eigenvalues of A, and δ is equal to the largest d for which (27) *has a solution with $\lambda = \tilde{E}$ or $-\tilde{E}$, there being only one such solution for $d = \delta$, i.e., $d(\tilde{E}) = d(-\tilde{E}) \pm 1$, and*

$$\delta = \max(d(\tilde{E}), d(-\tilde{E})). \tag{52}$$

Remarks. (1) The dual matrix B and function $g(x)$, having been introduced in order to prove the existence of the asymptotic solution, and to uncover its properties, have served their purpose: they are not needed for finding the solution to a specific problem.

(2) Bounds on δ are given by (44), with p and q the orders of the eigenvalues $\tilde{E}$ and $-\tilde{E}$.

(3) Just as for optimal approximation, our solution must give $\phi_0 \neq 0$, $\phi_s \neq 0$ and so normalizable to 1, and Q/D in lowest terms. If $\delta > 0$ we can proceed as if $\delta = 0$, using either $\lambda = \tilde{E}$ or $-\tilde{E}$, and will obtain an equivalent Q/D though with cancelling common factors.

(4) A theorem of Elliott and Lam [7, Theorem 4.2] states, in the notation of this paper: If $A\boldsymbol{\phi} = \lambda\boldsymbol{\phi}$, $|\lambda| =$ maximum eigenvalue modulus of A, $\phi_0 \neq 0$, and $D_1(\lambda) \neq 0$, $D_1(-\lambda) \neq 0$, then $\tilde{E} = |\lambda|$ and no other eigenvalue has this modulus.

In fact, when $D_1(\lambda) \neq 0$ (for *any* eigenvalue λ of A) the condition $\phi_0 \neq 0$ is superfluous, by Lemma 2 with $p = 0$. This means also that [7, Lemma 4.5], which gives a sufficient condition for $\phi_0 \neq 0$, is also superfluous. (Moreover in the condition given, namely, that $D_2(\lambda) \neq 0$, $D_2(-\lambda) \neq 0$, the second part is irrelevant, for, as the proof of our Lemma 2 shows, $D_2(\lambda) \neq 0 \Rightarrow \phi_0 \neq 0$.)

By our Theorem 2 it is always true that $\tilde{E} =$ maximum eigenvalue modulus of A. The conditions on D_1 in [7, Theorem 4.2] merely ensure the uniqueness of λ, for they imply $d(\lambda) = 0$, $d(-\lambda) \leqslant 0$ (in fact $d(-\lambda) = -1$, since $d(-\lambda) \neq d(\lambda)$ when $\lambda = \pm\tilde{E}$), and hence $\delta = 0$.

Example. As a simple illustration of a case with positive deficiency, and therefore not covered by [7, Theorem 4.2], consider the problem of approximating to $T_{n+3}(x)$ by a polynomial of degree n. (The solution is of course the zero polynomial, by the alternation theorem, for the error function has not merely the necessary $n + 2$ but in fact $n + 4$ alternation points.)

Here $r = 2$, $a_2 = 1$, $a_1 = a_0 = 0$. The eigenvalues of A are 1, 1, -1, so that $\hat{E} = \tilde{E} = 1$. For $\lambda = 1$ and $d = 0, 1, 2$ the first $s + 1 = 3 - d$ columns of $A - \lambda S^d$ have rank 1, 1, 0, i.e., in each case $\leqslant s$. Thus the largest d for which this is true is $d(1) = 2$. Similarly, for $\lambda = -1$ the ranks are 2, 1, 1, and $d(-1) = 1$. Thus the deficiency $\delta = 2 = d(1)$, and we must use $\lambda = 1$ in solving the problem. Then (27) gives $\boldsymbol{\phi} = [1]$, $\phi(u) = 1$, $M = -\frac{1}{2}(u^{n+3} + u^{-n-3}) = -T_{n+3}(x)$, $D = 1$, and hence $Q = M + Df = 0$, giving the zero polynomial as our solution. We note that the number of alternation points is indeed $n + 2 + \delta$ as predicted in Section 3.

Now A_1 has eigenvalues 1, -1 and A_2 has eigenvalue 1. Clearly the condition on A_1 in [7, Theorem 4.2] is not satisfied. Similarly, [7, Theorem 3.2], which requires (in our notation) $\phi_0 \neq 0$ and all roots of $\phi(u)$ outside the unit circle, where $\boldsymbol{\phi}$ is an eigenvector of A, is also inapplicable, for $\lambda = 1$ has general eigenvector $(1, c, 1)'$, with c arbitrary, and $\lambda = -1$ has eigenvector $(1, 0, -1)'$, and in both cases the condition on $\phi(u)$ is not satisfied.

A more substantial example is given in Appendix 2, which also describes a simple procedure for finding the polynomial $\mathrm{Int}(Q/D)$ without finding Q and D.

7. Clenshaw's Conjecture

As we have seen, the asymptotic approximation sought is given by the unique solution of (26), where $|\lambda| = \tilde{E}$, the maximum eigenvalue modulus of A, and $s = r - \delta$, δ being the deficiency of the problem. To the solution $\boldsymbol{\phi}$ of (26) corresponds a polynomial $\phi(u) = \phi_s u^s + \cdots + \phi_0$, with ϕ_s and $\phi_0 \neq 0$. Now let

$$\psi(u) = (u + 1)\, \phi(u) = \psi_{s+1} u^{s+1} + \cdots + \psi_0\,, \qquad \boldsymbol{\psi} = (\psi_0\,,\ldots,\, \psi_{s+1})'.$$

Then it is easy to verify that (as already indicated in Section 4)

$$A \boldsymbol{\psi}_{(\delta-1)} = \lambda S^{\delta-1} \boldsymbol{\psi}_{(\delta-1)}\,.$$

By repeated multiplication by $u + 1$ it is clear we eventually obtain

$$\phi^*(u) = (u + 1)^\delta\, \phi(u) = \phi_r^* u^r + \cdots + \phi_0^*, \qquad \phi_r^* = \phi_s \neq 0,$$

and

$$A\boldsymbol{\phi}^* = \lambda\boldsymbol{\phi}^*.$$

We have thus established that to the eigenvalue λ of A corresponds an eigenvector $(\phi_0^*,\ldots,\, \phi_r^*)$ with ϕ_r^* (and hence ϕ_0^*) $\neq 0$. For simplicity we shall now drop the asterisk, and normalize $\boldsymbol{\phi}$ by taking $\phi_0 = 1$:

$$A(1,\, \phi_1\,,\ldots,\, \phi_r)' = \lambda(1,\, \phi_1\,,\ldots,\, \phi_r)', \tag{53}$$

with

$$\lambda = \epsilon \tilde{E}, \qquad \epsilon = \pm 1.$$

Now Clenshaw in [6] was interested in finding the maximum ratio of the error norms S_n and E_n, given in (2) and (3), for all possible given polynomials $f(x)$, i.e., all possible coefficients $a_r\,,\ldots,\, a_0$. It is of course difficult to compute S_n for given a's but we shall, following Clenshaw (who confirmed this empirically in a number of cases) make the plausible assumption that when S_n/E_n is maximum, the norm S_n is attained at $x = \pm 1$, i.e., the a's are either all of the same sign, or of alternating signs. The latter case becomes the former on changing x to $-x$, so without loss of generality we shall assume the a's all of the same sign, and

$$S_n = | a_r + \cdots + a_0 | . \tag{54}$$

Further, we know from (25) that $\tilde{E}/E_n \to 1$ as $n \to \infty$ (with r and the a_i fixed). Thus, letting

$$\rho = (a_r + \cdots + a_0)/\tilde{E}, \tag{55}$$

we shall choose the a_i so as to maximize $| \rho |$. If we normalize the a_i by writing

$$c_i = a_i/\epsilon\tilde{E}, \qquad i = 0,\ldots, r. \tag{56}$$

then (53) becomes, on rearrangement,

$$\begin{bmatrix} 1 & \phi_1 - 1 & \cdot & \cdot & \cdot & \phi_r - 1 \\ & 1 & \phi_1 & & & \phi_{r-1} \\ & & \cdot & & & \cdot \\ & & & \cdot & & \cdot \\ & 0 & & & \cdot & \phi_1 \\ & & & & & 1 \end{bmatrix} \begin{bmatrix} \epsilon\rho \\ c_1 \\ \cdot \\ \cdot \\ c_{r-1} \\ c_r \end{bmatrix} = \begin{bmatrix} 1 \\ \phi_1 \\ \cdot \\ \cdot \\ \cdot \\ \phi_r \end{bmatrix} . \tag{57}$$

We may now solve for ρ and obtain

$$\rho = \epsilon F_r , \; F_r = \begin{vmatrix} 1 & \phi_1 - 1 & \cdot & \cdot & \cdot & \phi_r - 1 \\ \phi_1 & 1 & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \cdot & & 0 & & \cdot & \cdot \\ \phi_r & & & & & 1 \end{vmatrix} = F_{r-1} + \phi_r H_r , \tag{58}$$

where $F_0 = 1$, $H_1 = 1 - \phi_1$, and

$$\begin{aligned} H_r = (-1)^r & \begin{vmatrix} \phi_1 - 1 & \cdot & \cdot & \cdot & \phi_r - 1 \\ 1 & \phi_1 & & & \phi_{r-1} \\ & \cdot & & & \cdot \\ & 0 & \cdot & & \cdot \\ & & & 1 & \phi_1 \end{vmatrix} \\ & = 1 - \phi_r - \phi_{r-1} H_1 - \cdots - \phi_1 H_{r-1} . \end{aligned} \tag{59}$$

Clearly, $\partial H_r/\partial \phi_r = -1$, $\partial H_r/\partial \phi_t = 0$, $t > r$. For $t < r$ we have:

LEMMA 5. *$\partial H_r/\partial\phi_t = \sum_{k=1}^{r-t} \gamma_{r-t-k}(\phi_k - H_k)$, $1 \leqslant t < r$, where*

$$\gamma_0 = 1, \qquad \gamma_p = -\sum_{q=1}^{p} \gamma_{p-q}\phi_q, \qquad 0 < p \leqslant r - t - 1. \tag{60}$$

Proof. We use induction on r. First, $H_2 = \phi_1^2 - \phi_1 - \phi_2 + 1$, and $\partial H_2/\partial\phi_1 = 2\phi_1 - 1 = \phi_1 - H_1$, so that the lemma holds for $r = 2$. Now suppose it holds for $r = 2, 3,..., r - 1$. Then for $t < r$,

$$\begin{aligned}
\partial H_r/\partial\phi_t &= -H_{r-t} - \sum_{q=1}^{r-1} \phi_q(\partial H_{r-q}/\partial\phi_t) \\
&= -H_{r-t} + \phi_{r-t} - \sum_{q=1}^{r-t-1} \phi_q \left[\sum_{k=1}^{r-q-t} \gamma_{r-q-t-k}(\phi_k - H_k)\right] \\
&= (\phi_{r-t} - H_{r-t}) - \sum_{k=1}^{r-t-1} (\phi_k - H_k)\left[\sum_{q=1}^{r-t-k} \gamma_{(r-t-k)-q}\phi_q\right] \\
&= \sum_{k=1}^{r-t} \gamma_{r-t-k}(\phi_k - H_k), \quad \gamma_{r-t-1} = -\sum_{q=1}^{r-t-1} \gamma_{r-t-1-q}\phi_q\,,
\end{aligned}$$

which proves the lemma.

Now by (58),

$$F_r = 1 + \sum_{p=1}^{r} \phi_p H_p\,. \tag{61}$$

Hence

$$\begin{aligned}
&\partial F_r/\partial\phi_r = H_r - \phi_r\,, \\
&\partial F_r/\partial\phi_t = H_t - \phi_t + \sum_{p=t+1}^{r} \phi_p\left[\sum_{k=1}^{p-t} \gamma_{p-t-k}(\phi_k - H_k)\right], \qquad t < r.
\end{aligned}$$

Thus a sufficient condition that ρ is a stationary function of $\phi_1,..., \phi_r$ is that

$$\phi_k = H_k\,, \qquad k = 1,..., r, \tag{62}$$

i.e.,

$$\sum_{0}^{r} \phi_t\phi_{r-t} = 1.$$

This means that for small u,

$$\begin{aligned}
\phi^2(u) &= 1 + u + \cdots + u^r + O(u^{r+1}) \\
&= (1 - u)^{-1} + O(u^{r+1})
\end{aligned}$$

whence

$$\phi(u) = (1 - u)^{-1/2} + O(u^{r+1}). \tag{63}$$

In other words,

$$\begin{aligned}\phi_k &= \text{coefficient of } u^k \text{ in } (1 - u)^{-1/2}, \qquad k = 1, 2, \ldots \\ &= \frac{1.\,3.\,\cdots\,(2k - 1)}{2.\,4.\,\cdots\,2k},\end{aligned} \tag{64}$$

which was Clenshaw's conjecture.

The corresponding value of $S_n/\tilde{E}$, i.e., $|\,\rho\,|$, is then, by (58), (61) and (62),

$$F_r = 1 + \sum_1^r \phi_p^2. \tag{65}$$

The values of the γ's in (60) are easily determined by writing

$$\gamma(u) = 1 + \gamma_1 u + \cdots + \gamma_r u^r.$$

It then follows by (60) that $\gamma(u)\ \phi(u) = 1 + O(u^{r+1})$, whence

$$\gamma(u) = (1 - u)^{1/2} + O(u^{r+1}),$$

and

$$\begin{aligned}\gamma_k &= \text{coefficient of } u^k \text{ in } (1 - u)^{1/2}, \qquad k = 1, 2, \ldots \\ &= \frac{-1.\,1.\,3.\,\cdots\,(2k - 3)}{2.\,4.\,6.\,\cdots\,2k}.\end{aligned} \tag{66}$$

Further, if C denotes the matrix A/λ, we have

$$C\boldsymbol{\phi} = \boldsymbol{\phi}, \tag{67}$$

which gives the c's in succession from $\boldsymbol{\phi}$ by

$$\begin{aligned}c_r &= \phi_r \\ c_{r-1} &= \phi_{r-1} - c_r\phi_1 \\ c_{r-2} &= \phi_{r-2} - c_{r-1}\phi_1 - c_r\phi_2 \\ &\cdots\cdots\cdots\cdots \\ c_0 &= 1 - c_1\phi_1 - c_2\phi_2 - \cdots - c_r\phi_r\,.\end{aligned} \tag{68}$$

It is not immediately apparent that if the ϕ's of (64) are substituted here, the resulting c's are all of the same sign (i.e., positive), without which (54) and hence our whole solution is invalid. However, it can be shown that the values $c_i^{(r)}$ of c_i corresponding to any value of r are given by

$$c_i^{(r)} = \frac{2r+1}{2i+1}\, \phi_{r-i}\phi_r\,, \qquad i = 0, 1,\dots, r, \tag{69}$$

and thus are all positive as required. A proof of (69) is given in Appendix 1.

What we have shown, then, is that the ϕ's of (64) give a matrix C with positive elements c_i and eigenvalue unity (or equivalently a matrix $A = \lambda C$ with elements a_i all of the same sign and eigenvalue λ—λ being an arbitrary scaling factor), and are such as to make the corresponding sum $c_0 + \cdots + c_r$ (i.e., $(a_0 + \cdots + a_r)/\lambda$) a stationary function of the ϕ's. We have not however shown that the eigenvalue 1 is an eigenvalue of maximum modulus for C, nor that the stationary function is in fact a global or even a local maximum. Clenshaw [6] verified the global maximum property in the cases $r = 1, 2$, and 3, and Lam and Elliott [8] reported that they had verified the local maximum property in the cases $r = 1, 2, 3$, and 4. The global maximum property for general r remains unproved, and at present I see no way of proving it.

On the other hand the maximum modulus property for the eigenvalue 1 of C, or λ of A, is equivalent, as we have seen, to the polynomial $\phi(u)$ having no roots inside the unit circle. Thus to prove it we must prove that all partial sums $1 + \frac{1}{2}u + \cdots$ of the Maclaurin series for $(1 - u)^{-1/2}$ have no roots inside the unit circle. This follows, as the coefficients are nonincreasing and positive, by the Eneström–Kakeya Theorem (see, for example, [12]).

Assuming therefore that

(a) when S_n/E_n is maximum, S_n is attained at $x = \pm 1$, and

(b) $c_0 + \cdots + c_r$ is maximum when the ϕ_k are as in (64),

we have shown that for all $f(x)$ as in (1), and large n,

$$S_n/E_n \sim S_n/\tilde{E} \leqslant 1 + \sum_{1}^{r} \phi_k^2.$$

Appendix 1

Proof of (69): Since (68) determines the $c_i^{(r)}$ uniquely for a particular r, we can prove (69) by showing that the values of $c_i = c_i^{(r)}$ in (69) satisfy (68), i.e.

$$(2r+1)\, \phi_r \sigma_{r,k} = \phi_k\,, \qquad k = 0,\dots, r \tag{A1}$$

where

$$\sigma_{r,k} = \frac{\phi_0\phi_{r-k}}{2k+1} + \frac{\phi_1\phi_{r-k-1}}{2k+3} + \cdots + \frac{\phi_{r-k}\phi_0}{2r+1}$$

$$= \text{coefficient of } u^{2r+1} \text{ in } (1-u^2)^{-1/2}\int_0^u u^{2k}(1-u^2)^{-1/2}\,du.$$

Differentiating

$$(1-u^2)^{-1/2}\int_0^u u^{2k}(1-u^2)^{-1/2}\,du = \sum_k^\infty \sigma_{r,k}u^{2r+1}$$

gives

$$u(1-u^2)^{-3/2}\int_0^u u^{2k}(1-u^2)^{-1/2}\,du + (1-u^2)^{-1}\,u^{2k} = \sum_k^\infty (2r+1)\,\sigma_{r,k}u^{2r}$$

or

$$u\sum_k^\infty \sigma_{r,k}u^{2r+1} + u^{2k} = \sum_k^\infty (1-u^2)(2r+1)\,\sigma_{r,k}u^{2r}.$$

Equating coefficients of u^{2r}, $r > k$ gives the recursive relation

$$\sigma_{r,k} = (2r/(2r+1))\,\sigma_{r-1,k}\,,$$

whence (A1) follows.

Appendix 2

Practical Considerations and Example

For a given $f(x)$, both the optimal and the asymptotic Q/D can be found as described in Section 4 with appropriate choice of λ. In the asymptotic case, once λ and $\phi(u)$ have been found, it is easy to determine the required integral part of Q/D without actually finding Q and D. We have by (14) and (20)

$$\frac{M(x)}{D(x)} = \frac{1}{2}(\sigma(u) + \sigma(u^{-1})), \qquad \sigma(u) = -\lambda u^{n+1+d}\,\frac{\phi(u)}{\phi(u^{-1})}\,. \tag{A2}$$

Now if we write

$$\sigma(u) = \sigma_m u^m + \cdots + \sigma_0 + \text{fractional part}, \tag{A3}$$

then

$$\sigma_i = -a_{i-n-1}\,, \qquad i = m, m-1, \ldots, n+1, \tag{A4}$$

and $\sigma_n ,\ldots, \sigma_0$ can be found successively from

$$\phi_0\sigma_i + \phi_1\sigma_{i+1} + \cdots + \phi_s\sigma_{i+s} = 0, \qquad i = n, n-1,\ldots, 0. \tag{A5}$$

It is then easy to see, since $\sigma(0) = 0$, that the fractional part of $\sigma(u)$ contributes $-\frac{1}{2}\sigma_0$ to the integral part of M/D, and hence that, apart from the "lower order terms" in (1),

$$P = \operatorname{Int}(Q/D) = \sigma_n T_n(x) + \cdots + \sigma_1 T_1(x) + (1/2)\,\sigma_0 . \tag{A6}$$

As an illustrative example for the whole solution procedure, let

$$f(x) = T_{n+4} + (1/2)\,T_{n+3} + (5/4)\,T_{n+2} - (7/8)\,T_{n+1}\,,$$

and suppose an asymptotic solution is required. The matrix

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

has eigenvalues $\lambda = 2,\ -2,\ (-3 \pm 73^{1/2})/16$. Thus $\tilde{E} = 2$, and by (44) $\delta = 1$. Since

$$\begin{bmatrix} -7/8 & 5/4 & 1/2 \\ -3/4 & 1/2 & 1 \\ 1/2 & -1 & \\ 1 & & -2 \end{bmatrix}$$

has rank 2, with column-multipliers 2, 1, 1, we must have $d(2) = \delta$ and $d(-2) = 0$ (which are easily confirmed), and $s = 2$. Also

$$\phi = (2, 1, 1)', \qquad \phi(u) = 2 + u + u^2,$$

with roots of modulus $2^{1/2}$, i.e., greater than 1 as expected.

We can now proceed at once to find the polynomial approximation to f. By (A4),

$$\sigma_{n+1} = 7/8, \quad \sigma_{n+2} = -5/4, \quad \sigma_{n+3} = -1/2, \quad \sigma_{n+4} = -1,$$

and by (A5) $\sigma_i = -\frac{1}{2}(\sigma_{i+1} + \sigma_{i+2})$, $i \leqslant n$, giving

$$\sigma_n = 3/16, \quad \sigma_{n-1} = -17/32, \quad \sigma_{n-2} = 11/64, \quad \sigma_{n-3} = 23/128,\ldots.$$

If for definiteness we take $n = 3$, then

$$f = T_7 + (1/2)\,T_6 + (5/4)\,T_5 - (7/8)\,T_4\,,$$

and our approximation to it of degree 3 is

$$P = (3/16)\,T_3 - (17/32)\,T_2 + (11/64)\,T_1 + (23/256).$$

Alternatively, P can be found by using $\lambda = 2$ or $\lambda = -2$ and $d = 0$. With $\lambda = 2$, ϕ is the normalized eigenvector $(2, 3, 2, 1)'$ and

$$\phi(u) = 2 + 3u + 2u^2 + u^3 = (1 + u)(2 + u + u^2).$$

Taking $\sigma_{n+1}, \ldots, \sigma_{n+4}$ as above, σ_i are now found from

$$\sigma_i = -\tfrac{1}{2}(3\sigma_{i+1} + 2\sigma_{i+2} + \sigma_{i+3}), \qquad i \leqslant n,$$

which give the same σ_i and P as before. Q and D have the common factor $x + 1$.

Similarly, with $\lambda = -2$, ϕ is the eigenvector $(-2, 1, 0, 1)'$,

$$\phi(u) = -2 + u + u^3 = (-1 + u)(2 + u + u^2),$$

again giving the same solution, and the common factor $x - 1$.

It may be of interest to compare the norm of error of P with the optimum error for polynomials of degree $n = 3$, making use of (25). Using the lowest-degree solution, we find

$$D = x^2 + \tfrac{3}{4}x + \tfrac{1}{4}$$

and

$$\omega_+ = 2[2(x + 1)]^{1/2} \cdot (4x^4 - 2x^2 - x),$$
$$\omega_- = [2(x - 1)]^{1/2} \cdot (8x^4 + 8x^3 - 2x - 1).$$

The roots of $R - 2$ in $[-1, 1]$ are $-1, 0, 0.885$, and of $R + 2$ are -0.858, 0.554, 1. For the error of P we have

$$R^* = P - f = R - \mathrm{Fr}\left(\frac{Q}{D}\right) = R + \frac{89x - 1}{1024D}.$$

Thus at the norm-points $-1, \ldots, 1$ of R, R^* takes successive values 1.824, -2.221, 1.996, -1.950, 2.045, -1.957, and bounds on E_3 are given by

$$1.950 \leqslant E_3 \leqslant \| R^* \| \doteq 2.221.$$

(In finding the lower bound we may ignore the error 1.824 at -1, since $5 = n + 2$ alternating errors remain.) Closer bounds (and an improved approximation) are obtained on replacing P by $P + 0.023$, which gives

$$1.973 \leqslant E_3 \leqslant 2.198.$$

Thus our polynomial P, with error norm 2.221, is a fair approximation already, and certainly a good starting point for an approximation algorithm. In this example n has the low value of 3. The goodness of approximation of P of course increases with n.

References

1. N. I. Achieser, "Theory of Approximation," Ungar, New York, 1956.
2. N. I. Achieser, Über die Funktionen die in gegebenen Intervallen am wenigsten von Null abweichen, *Izv. Kazan. Fiz. Mat.* **3** (1928).
3. S. N. Bernstein, "Leçons sur les propriétés extrémales et la meilleure approximation des fonctions analytiques d'une variable réelle," Gauthier–Villars, Paris, 1926.
4. P. L. Chebyshev, Théorie des mécanismes connus sur le nom de parallélogrammes, *in* "Oevres," Vol. I, pp. 111–143. Chelsea, New York.
5. E. W. Cheney, "Introduction to Approximation Theory," McGraw–Hill, New York, 1966.
6. C. W. Clenshaw, A comparison of "best" polynomial approximations with truncated Chebyshev series expansions. *SIAM J. Numer. Anal.* **1** (1964), 26–37.
7. D. Elliott and B. Lam, An estimate of $E_n(f)$ for large n. *SIAM J. Numer. Anal.* **10** (1973), 1091–1102.
8. B. Lam and D. Elliott, On a conjecture of C. W. Clenshaw. *SIAM J. Numer. Anal.* **9** (1972), 44–52.
9. T. J. Rivlin, "An Introduction to the Approximation of Functions," Blaisdell, Waltham, 1969.
10. A. Talbot, On a class of Tchebyscheffian approximation problems solvable algebraically. *Proc. Cambridge Philos. Soc.* **58** (1962), 244–267.
11. A. Talbot, The Tchebyscheffian approximation of one rational function by another. *Proc. Cambridge Philos. Soc.* **60** (1964), 877–890).
12. M. Marden, "Geometry of Polynomials," p. 136. American Mathematical Society, Providence, Rhode Island, 1966.