Journal of Scientific Computing, Vol. 5, No. 4, 1990

The Envelope of the Error for Trigonometric and Chebyshev Interpolation

John P. Boyd[1]

Received November 19, 1990

The error in Chebyshev or Fourier interpolation is the product of a rapidly varying factor with a slowly varying modulation. This modulation is the "envelope" of the error. Because this slow modulation controls the amplitude of the error, it is crucial to understand this "error envelope." In this article, we show that the envelope varies strongly with x, but its variations can be predicted from the convergence-limiting singularities of the interpolated function $f(x)$. In turn, this knowledge can be translated into a simple spectral correction algorithm for wringing more accuracy out of the same pseudospectral calculation of the solution to a differential equation.

KEY WORDS: Interpolation error; Fourier method; Chebyshev method; pseudospectral method.

1. INTRODUCTION

The N-point interpolant of a function $f(x)$ is defined to be that polynomial of degree $(N-1)$ that agrees with $f(x)$ at N interpolation points. This implies that the interpolation error $E_I(x; N)$ is zero at each of the N interpolation points. If we divide the error by a rapidly oscillating factor (defined more precisely below) which vanishes at these same N points, we obtain the slowly varying "error envelope" which is our main theme.

Figure 1 illustrates the exact interpolation error and its envelope for a typical function.

To illustrate the phenomenology of the error envelope, we shall concentrate on trigonometric cosine interpolation, choosing the interpolation

[1] Department of Atmospheric, Oceanic & Space Science, Laboratory for Scientific Computation, University of Michigan, 2200 Bonisteel Boulevard, Ann Arbor, Michigan 48109.

points to be the roots of $\cos(Nx)$. Because the Chebyshev polynomials are related to the cosines via the familiar identity

$$T_n(\cos x) = \cos(nx) \tag{1.1}$$

it follows that every example is simultaneously a Fourier interpolant and also the interpolant of $f(\arccos x)$ by an ordinary polynomial. The Chebyshev interpolation points are the roots of $T_N(y)$, which form the so-called "roots" or "Gauss–Chebyshev" grid.

For both trigonometric and polynomial interpolation, there is a second canonical grid: the so-called "Gauss–Lobatto" or "extrema-and-end

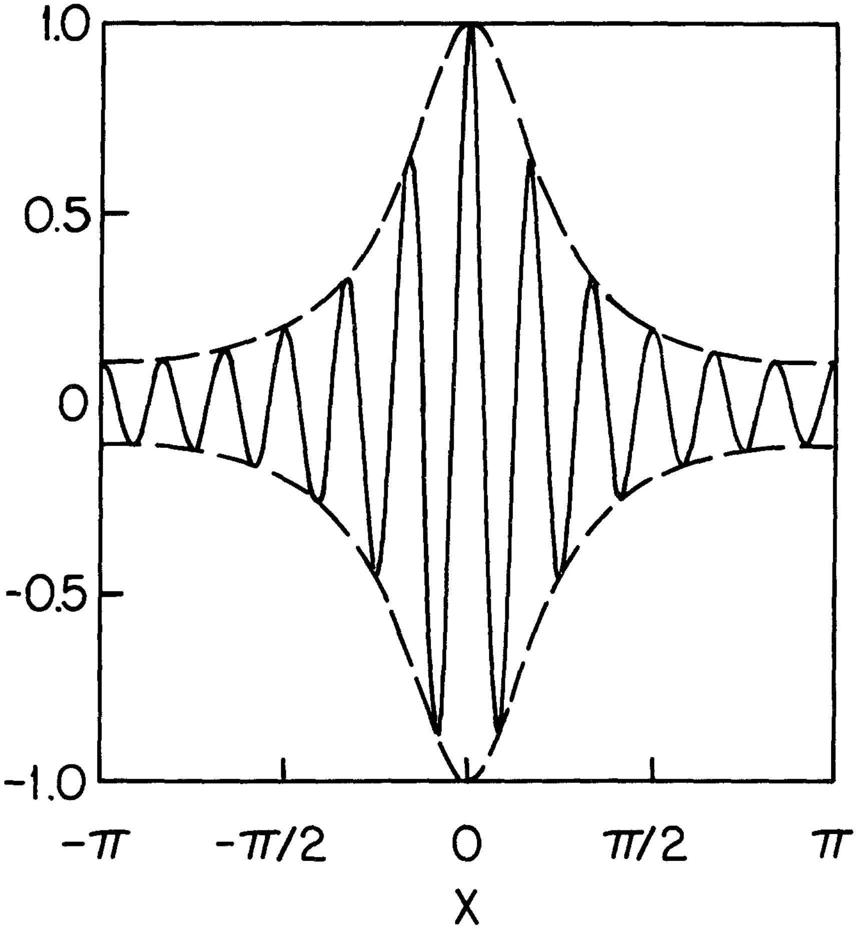

Fig. 1. Solid curve: the interpolation error $E_I(x; N = 12)$ for the model function $\lambda(x, p = 0.5)$, which is defined by (2.24). Dashed curve: the envelope of the interpolation error ρ (and its reflection with respect to the x axis).

points" grid. This alternative set of points is especially popular for solving differential equations with Dirichlet boundary conditions. However, the two canonical interpolation grids are so closely related that it would be foolish and redundant to quote every theorem twice. Instead, we shall limit ourselves to the "roots" grid because it is the simpler.

We begin with a few essential definitions which the spectral maven should read at near-light speed. The N-point interpolant of $f(x)$ is

$$I_N f(x) = a_0/2 + \sum_{j=1}^{N-1} a_j \cos(jx) \tag{1.2}$$

where the coefficients $\{a_j\}$ are determined by the N interpolation conditions (1.3)

$$I_N f(x_i) = f(x_i) \tag{1.3}$$

where

$$x_i = (2i-1)\pi/(2N) \leftrightarrow \cos(Nx_i) = 0 \tag{1.4}$$

The N-term projection of $f(x)$, also known as the N-term truncation of the series for $f(x)$, is defined by

$$P_N f(x) = \alpha_0/2 + \sum_{j=1}^{N-1} \alpha_j \cos(jx) \tag{1.5}$$

where the $\{\alpha_j\}$ are the *exact* special coefficients defined by

$$\alpha_j \equiv (2/\pi) \int_0^\pi f(x) \cos(jx)\, dx \tag{1.6}$$

The interpolation error $E_I(x; N)$ and the truncation error $E_T(x; N)$ are then defined by

$$E_I(x; N) \equiv |f(x) - I_N f(x)| \tag{1.7a}$$

$$E_T(x; N) \equiv |f(x) - P_N f(x)| \tag{1.7b}$$

Definition 1. The *envelope* of the *interpolation error* is the function $\rho(x; N)$ in the factorization of the interpolation error:

$$E_I(x; N) = \cos(Nx)\, \rho(x; N) \tag{1.8}$$

i.e.,

$$\rho(x; N) \equiv E_I(x; N)/\cos(Nx) \tag{1.9}$$

By its very definition, the interpolation error $E_I(x; N)$ is zero at the roots of $\cos(Nx)$, so the ratio $E_I/\cos(Nx)$ is always nonsingular except at singularities of $f(x)$.

The $\cos(Nx)$ factor in (1.8) implies that the interpolation error will oscillate very rapidly. The information content of these oscillations is nil, however, because all the troughs and crests of $\cos(Nx)$ are identical. The *magnitude* and *spatial uniformity* of the error are *controlled* entirely by the *envelope*. It follows that it is only common sense to focus upon the envelope instead of $E_I(x; N)$.

We will pursue a fourfold strategy to analyze the envelope. In the next section, we prove exact theorems that relate the spectral coefficients of the interpolation and truncation errors to those of $f(x)$. We also derive closed-form errors for two simple model functions, $\lambda(x; p)$ and $\mu(x; p)$.

Section 3 develops the simple but very accurate "Neglect-of-Triple-Aliasing" or "NTA" approximation. Under very general circumstances (and additional approximations), the envelope is proportional to some linear combination of the functions $\lambda(x; p)$ and $\mu(x; p)$.

Section 4 shows that the coefficients of the interpolation theory are "paired" in the sense that each coefficient has a twin of different degree but approximately the same magnitude. Section 5 illustrates the exact and approximate error envelopes for a variety of model functions: functions with poles, functions with logarithms, functions with nonanalytic but infinitely differentiable singularities. These representative cases give good insight into the possibilities. Lest this "model function" approach be taken too far, we offer a counterexample of a function with a *fractal* distribution of poles whose numerical behavior is not well understood.

The next topic is to use method-of-multiple-scales ideas to extend error envelope concepts to the pseudospectral solution of differential equations. The error in the solution to a differential equation usually does *not* vanish at the interpolation points. Nevertheless, we show that the error is *approximately* proportional to the residual of the differential equation. This in turn provides two rewards. The first is that one can estimate the error in the solution $u(x)$ by substituting the approximate numerical solution into the differential equation; the error is approximately this residual divided by N^2 (for a second-order equation).

The second reward is that one can correct the N-point pseudospectral solution to refine the answer to compute more than N coefficients *without* solving a matrix system of dimension greater than N (Sec. 7). This spectral correction algorithm, which gives an asymptotic series here, can be modified to give the convergent iterations described in Boyd (1991).

2. EXACT THEOREMS

Although our special interest is in Fourier cosine interpolation, which is equivalent under the change of variable $x = \cos(y)$ to Chebyshev interpolation, it is necessary for completeness to prove some theorems for sine interpolation, too. If the function $g(x)$ is antisymmetric about the origin, that is, $g(x) = -g(-x)$ for all x, then $g(x)$ can be represented as a sine series:

$$g(x) = \sum_{j=1}^{\infty} \beta_j \sin(jx) \tag{2.1}$$

The N-term projections and interpolants are written

$$P_N g(x) \equiv \sum_{j=1}^{N} \beta_j \sin(jx) \tag{2.2}$$

$$I_N g(x) \equiv \sum_{j=1}^{N} b_j \sin(jx) \tag{2.3}$$

Note that the sine sums have the range $(1, ..., N)$ versus the $(0, ..., N-1)$ range for their cosine counterparts. The corresponding truncation and interpolation errors and the envelope of the interpolation error are still defined by (1.7) and (1.9). However, we shall denote the sine error envelope by $\sigma(x; N)$ (versus $\rho(x; N)$ for the cosine error envelope).

Note that we lose no generality by considering only cosine and sine interpolation. An *arbitrary, nonsymmetric* function $u(x)$ can always be decomposed into parts that are symmetric and antisymmetric about $x = 0$ by writing

$$u(x) = f(x) + g(x) \tag{2.4}$$

where

$$f(x) \equiv [u(x) + u(-x)]/2 \qquad \text{(symmetric, cosine part)} \tag{2.5a}$$

$$g(x) \equiv [u(x) - u(-x)]/2 \qquad \text{(antisymmetric, sine part)} \tag{2.5b}$$

Thus, by evaluating the original function $u(x)$ at the roots of $\cos(Nx)$ on the interval $[-\pi, \pi]$, we can compute the values of $f(x)$ and $g(x)$ on the interval $[0, \pi]$. The problem of general Fourier interpolation on $x \in [-\pi, \pi]$ is split into two smaller problems on the half-interval, $x \in [0, \pi]$: cosine interpolation for $f(x)$ and sine interpolation for the antisymmetric component $g(x)$.

The following theorem is extremely important in characterizing the error envelope, as we shall see below.

Theorem 1: *High Degree Coefficients of the Truncation and Interpolation Errors.* Let $f(x)$ have the exact spectral series

$$f(x) = \alpha_0/2 + \sum_{j=1}^{\infty} \alpha_j \cos(jx) \tag{2.6}$$

Let the spectral coefficients of the truncation error and interpolation error be denoted by

$$E_T(x; N) = \sum_{j=N}^{\infty} e_j^{(T)} \cos(jx) \tag{2.7}$$

$$E_I(x; N) = e_0^{(I)}/2 + \sum_{j=1}^{\infty} e_j^{(I)} \cos(jx) \tag{2.8}$$

Then

$$\text{(i)} \qquad e_j^{(T)} = \alpha_j, \qquad j \geqslant N \tag{2.9}$$

$$\text{(ii)} \qquad e_j^{(I)} = \alpha_j, \qquad j \geqslant N \tag{2.10}$$

Equations (2.9) and (2.10) are also true for *sine* interpolation except that the lower index is $j = N + 1$ instead of $j = N$, consistent with the fact that $\sin[(N+1)x]$ is the first term omitted from (2.2) and (2.3).

The *low*-degree coefficients of the truncation and interpolation errors are given by Theorem 5 in Sec. 4.

Proof. The projection and the interpolant of $f(x)$ are both trigonometric polynomials of degree at most $(N-1)$. It follows that subtracting these polynomials from $f(x)$ to give $E_T(x; N)$ and $E_I(x; N)$ cannot alter or modify the high-degree coefficients of $f(x)$, which are passed on without modification to the differences, $E_T(x; N) \equiv |f(x) - P_N f(x)|$ and $E_I(x; N) \equiv |f(x) - I_N f(x)|$, as expressed by (2.9) and (2.10). □

Theorem 2a: *Fourier Coefficients of the Error Envelope (Cosine).* Let the interpolation error envelope, $\rho(x; N) \equiv |f(x) - I_N f(x)|/\cos(Nx)$, have the Fourier series

$$\rho(x; N) \equiv \rho_0/2 + \sum_{j=1}^{\infty} \rho_j \cos(jx) \tag{2.11}$$

Then these Fourier coefficients are related those of $f(x)$, the function being interpolated, by the difference equation

$$\rho_j + \rho_{j+2N} = 2\alpha_{j+N}, \qquad j = 0, 1,... \tag{2.12}$$

This difference equation has the exact infinite series solution

$$\rho_j = 2 \sum_{m=0}^{\infty} (-1)^m \alpha_{j+(2m+1)N}, \qquad j = 0, 1,... \tag{2.13}$$

Proof. The first step is to substitute the series (2.11) into the definition of $E_I(x; N)$ as the product of $\cos(Nx)$ with $\rho(x; N)$:

$$E_I(x; N) = (\rho_0/2)\cos(Nx) + \sum_{j=1}^{\infty} \rho_j \cos(jx)\cos(Nx) \tag{2.14}$$

Applying the standard trigonometric identity for the product of two cosines gives

$$E_I(x; N) = (\rho_0/2)\cos(Nx) + \sum_{j=1}^{\infty} \rho_j \{\cos[(N+j)x] + \cos[|N-j|\,x\}/2 \tag{2.15}$$

$$\begin{aligned} E_I(x; N) = (\rho_0/2)\cos(Nx) + &\sum_{k=N+1}^{\infty} \rho_{k-N}\cos(kx)/2 \\ &+ (1/2)\sum_{k=0}^{N-1} \rho_{N-k}\cos(kx) + (1/2)\sum_{k=1}^{\infty} \rho_{N+k}\cos(kx) \end{aligned} \tag{2.16}$$

The second step is to collect terms in (2.16) that are proportional to $\cos(kx)$, which gives

$$e_k^{(I)} = (1/2)(\rho_{|k-N|} + \rho_{k+N}), \qquad k = 0, 1, 2,... \tag{2.17}$$

This relation will be important in discussing the “coefficient pairing” in the interpolation error (Sec. 6).

The third step is to prove (2.12). We recall that Theorem 1 states that the high-degree coefficients of the error are identical with those of $f(x)$, restrict (2.17) to $k \geqslant N$, and then substitute $j = k - N$.

The fourth and final step is to prove that the infinite series (2.13) satisfies the difference equation (2.12). If we substitute the series into (2.12), we find

$$\rho_j + \rho_{j+2N} = 2\sum_{m=0}^{\infty} (-1)^m \alpha_{j+(2m+1)N} + 2\sum_{m=0}^{\infty} (-1)^m \alpha_{j+(2m+3)N} \tag{2.18}$$

If we separate the $m=0$ term in the first sum and then pair the $(k+1)$st term in the first sum with the kth term in the second sum, we obtain

$$\rho_j + \rho_{j+2N} = 2\alpha_{j+N} + 2 \sum_{m=0}^{\infty} (-1)^m(-1+1)\,\alpha_{j+(2m+3)N} \qquad (2.19)$$

$$= 2\alpha_{j+N} \qquad (2.12\text{bis}) \quad \square$$

Theorem 2b: *Fourier Sine Coefficients of the Error Envelope.* Let the interpolation error envelope, $\sigma(x; N) \equiv [g(x) - I_N g(x)]/\cos(Nx)$, have the Fourier series

$$\sigma(x; N) \equiv \sum_{j=1}^{\infty} \sigma_j \sin(jx) \qquad (2.20)$$

Then the Fourier coefficients are related to those of $g(x)$ by the difference equation

$$\sigma_j + \sigma_{j+2N} = 2\beta_{j+N}, \qquad j = 1, 2,... \qquad (2.21)$$

This difference equation has the exact solution

$$\sigma_j = 2 \sum_{m=0}^{\infty} (-1)^m \beta_{j+(2m+1)N}, \qquad j = 1, 2,... \qquad (2.22)$$

[Equations (2.21) and (2.22) are identical with their counterparts for cosine interpolation except that the lower limit is $j=1$ instead of $j=0$.]

Proof. The proof is identical to its cosine counterpart, so we omit details except to note for future reference that the relationship between the error coefficients and envelope coefficients is, for sine interpolation,

$$\begin{aligned} e_k^{(I)} &= \tfrac{1}{2}(\sigma_{k-N} + \sigma_{k+N}), && k \geqslant (N+1) \\ e_N^{(I)} &= \tfrac{1}{2}\sigma_{2N} && k = N \qquad (2.23) \\ e_k^{(I)} &= \tfrac{1}{2}(\sigma_{k+N} - \sigma_{N-k}), && k = 1,..., N-1 \qquad \square \end{aligned}$$

The last exact results we shall present are for two crucial model functions. The symmetric model is

$$\lambda(x; p) = 1 + 2 \sum_{j=1}^{\infty} p^j \cos(jx) \qquad (\text{“Lorentzian”}) \qquad (2.24\text{a})$$

$$= -2\log(p) \sum_{m=-\infty}^{\infty} 1/[\log^2(p) + (x - 2\pi m)^2] \qquad (2.24\text{b})$$

$$= (1-p^2)/[(1+p^2) - 2p\cos(x)] \qquad (2.24\text{c})$$

The first series representation shows that the coefficients of $\lambda(x;\, p)$ are powers of a constant $p<1$, i.e., are the terms of a geometric series. We shall see in the next section that the simple form of the coefficients of $\lambda(x;\, p)$ gives this function a special role in the theory of Chebyshev interpolation.

The partial fraction expansion, also known as the "imbricate" series representation, shows that $\lambda(x;\, p)$ has simple poles along the lines $\operatorname{Im}(x) = \pm\log(p)$. As shown in more detail in Boyd (1989a) and many other sources, the spectral coefficients are always the signature of the function being expanded. We will develop this theme through many examples in Sec. 5.

The antisymmetric model is

$$\mu(x;\, p) = 2 \sum_{j=1}^{\infty} p^j \sin(jx) \qquad (\text{"Serpentine"}) \tag{2.25a}$$

$$= 2 \sum_{m=-\infty}^{\infty} (x - 2\pi m)/[\log^2(p) + (x - 2\pi m)^2] \tag{2.25b}$$

$$= 2p \sin(x)/[(1+p^2) - 2p\cos(x)] \tag{2.25c}$$

This function is merely $\lambda(x;\, p)$ multiplied by $\sin(x)$ and a constant.

These model functions are periodic generalizations ("imbrications") of the classical functions known as the "witch of Agnesi" and "Newton's serpentine" (Rucker, 1987). The witch of Agnesi has also become known as the Lorentzian function because Lorentz showed that this function is a good model for certain spectral lines. Strictly, we should refer to $\lambda(x;\, p)$ and $\mu(x;\, p)$ as the "periodic Lorentzian" and "periodic serpentine," but we shall usually omit the adjective "periodic."

Figure 2 illustrates $\lambda(x;\, p)$ and $\mu(x;\, p)$ for a particular value of the parameter p. Figure 3 is a pair of contour plots that show how these functions vary with both x and p. In particular, note that $\lambda(x;\, p)$ has a single peak, symmetric about $x=0$, which becomes taller and narrower as p increases. In the limit $p \to 1$, $\lambda(x;\, p)$ tends to a delta function. Similarly, $\mu(x;\, p)$ has a shorter, wider peak and a matching trough. For small positive p, $\mu(x;\, p)$ is the sine function. As p increases, the crest and trough become narrower, taller, and move closer to the origin.

Theorem 3: *Exact Interpolation and Truncation Errors for the Model Functions $\lambda(x;\, p)$ and $\mu(x;\, p)$.* For the functions defined by (2.24) and (2.25), the interpolation errors are—without approximation—

$$\lambda(x;\, p): \quad E_I(x;\, N) = [2p^N/(1+p^{2N})] \cos(Nx)\, \lambda(x;\, p) \tag{2.26a}$$

$$\mu(x;\, p): \quad E_I(x;\, N) = [2p^N/(1+p^{2N})] \cos(Nx)\, \mu(x;\, p) \tag{2.26b}$$

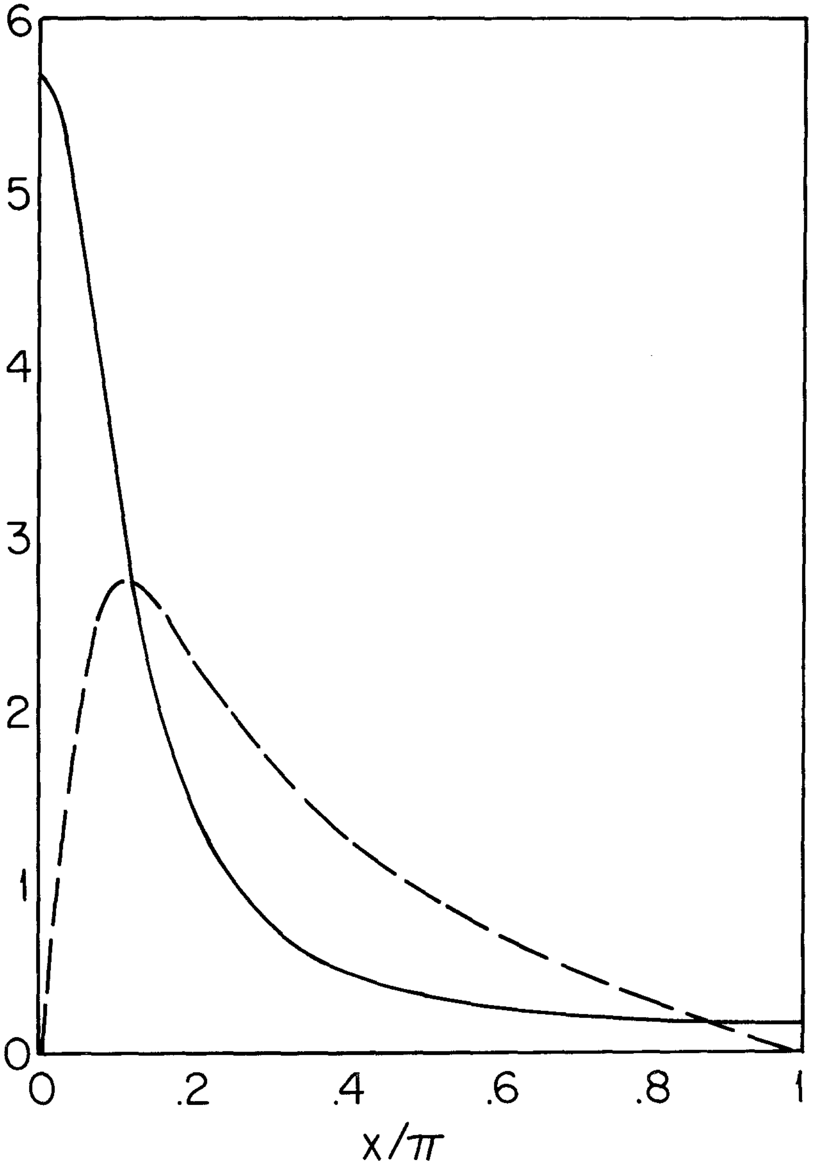

Fig. 2. Solid curve: $\lambda(x;\, p=0.7)$. Dashed curve: $\mu(x;\, p=0.7)$.

The corresponding exact truncation errors are

$\lambda(x; p)$:

$$E_T(x; N) = [2p^N/(1-p^2)]\{\cos(Nx) - p\cos[(N-1)x]\}\,\lambda(x; p) \quad (2.27a)$$

$\mu(x; p)$:

$$E_T(x; N) = -2p^{N+1}\{p\sin(Nx) - \sin[(N+1)x]\}/[(1+p^2) - 2p\cos(x)] \quad (2.27b)$$

(a)

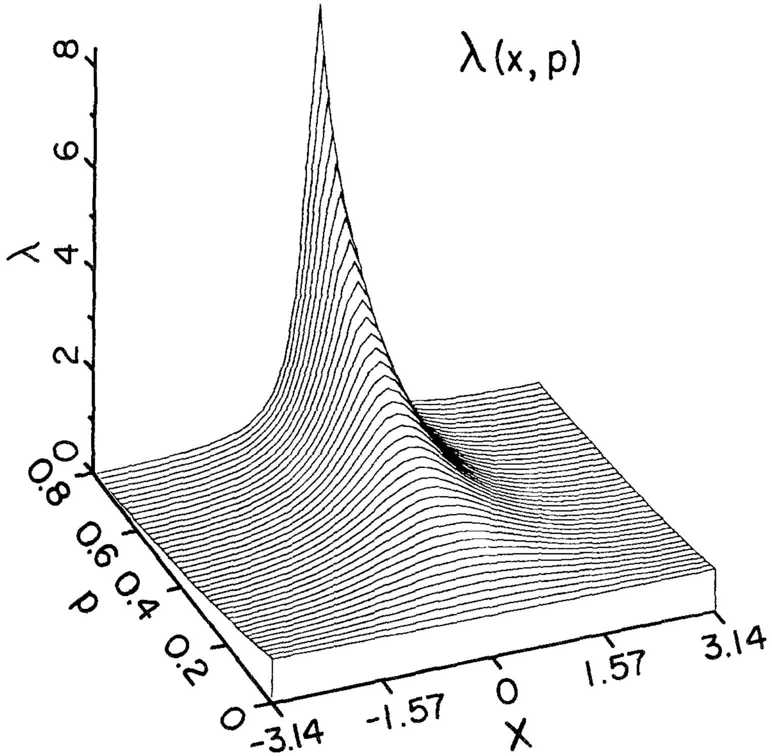

Fig. 3. (a) Contour plot of the periodic Lorentzian function $\lambda(x; p)$ for $p \in [0, 0.8]$. The peak becomes infinitely tall and narrow as $p \to 1$, so the plot is restricted to $p \leqslant 0.8$. In the limit $p \to 0$, $\lambda(x; 0)$ tends to a constant. (b) Contour plot of the periodic serpentine function $\mu(x; p)$ for $p \in [0, 0.8]$. The peak and trough become infinitely tall and narrow as $p \to 1$, so as in Fig. 3a, the range of p is restricted to $p \leqslant 0.8$. In the limit $p \to 0$, $\mu(x; p)$ is proportional to $\sin(x)$.

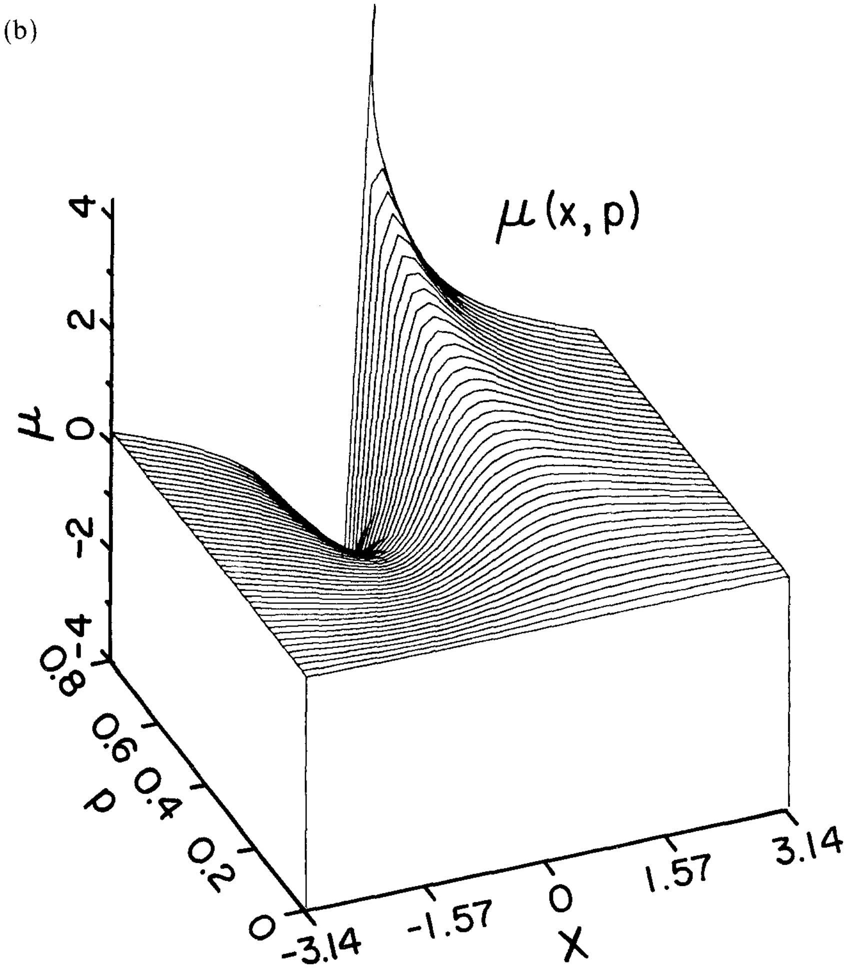

Fig. 3. Continued.

[Equations (2.27a) and (2.27b) were proved in Elliott (1965), through a contour integral representation.]

Proof. If we substitute $\alpha_j = 2p^j$ into Theorem 2, Eq. (2.13), we find that the coefficients of the interpolation error envelope are

$$\rho_j = 2(2p^N) \sum_{m=0}^{\infty} (-1)^m p^{j+2Nm} \tag{2.28}$$

$$= 2p^j(2p^N) \sum_{m=0}^{\infty} (-1)^m (p^{2N})^m \tag{2.29}$$

$$= 2p^j[2p^N/(1+p^{2N})] \tag{2.30}$$

where we identified the infinite sum in (2.29) as the geometric series for $1/(1+p^{2N})$. It follows that, except for the x-independent factor shown in square brackets in (2.30), the Fourier coefficients of the envelope are identical with those of $\lambda(x; p)$. Multiplying this envelope by $\cos(Nx)$ gives (2.26a). The proof of (2.26b) is so similar that we omit details. □

Gradshteyn and Ryzhik (1965) give the partial sums of the series for $\lambda(x; p)$ and $\mu(x; p)$ through the first N terms as their identities 1.353.1 and 1.353.2. Subtracting these analytic expressions from $\lambda(x; p)$ and $\mu(x; p)$ then gives formulas for the sums from N to ∞, i.e., for the truncation errors, (2.27).

Figure 4 compares the exact truncation error for $\lambda(x; p)$ with the

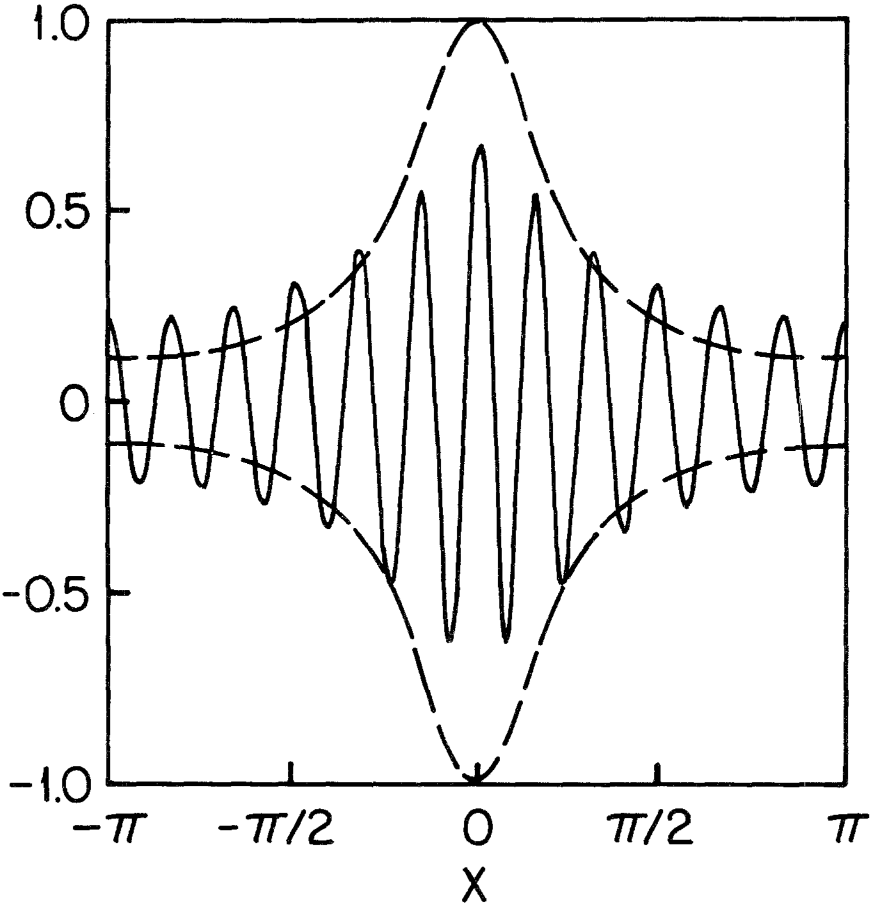

Fig. 4. A comparison of the truncation error $E_T(x;\ N=12)$ for $\lambda(x,\ p=0.5)$ (solid) with the envelope $\rho(x; N)$ of the interpolation error and the reflection of ρ (dashed).

interpolation error for the same function. One striking feature is that the *maximum* truncation error is *smaller* than the maximum interpolation error by almost a factor of 2. We shall see in Sec. 4 that this factor of 2 is an upper bound on $\max|E_I|/\max|E_T|$ for fixed N. The second striking feature is that the truncation error is more uniform than the interpolation error so that the interpolant is actually more accurate than the truncation of the spectral series near $x = \pm\pi$.

These theorems are also somewhat disappointing. The error for Lagrangian interpolation is the product of two factors: one that depends only on the function $f(x)$ that is being interpolated, and the other, which depends on the choice of interpolation points. The trigonometric (and Chebyshev polynomial) grids make this second factor as *uniform* as possible with respect to x.

Unfortunately, the $f(x)$-dependent factor is anything but uniform. The error in interpolating $\lambda(x; p)$ and $\mu(x; p)$ is directly proportional to $\lambda(x; p)$ and $\mu(x; p)$ themselves, respectively. Because these functions are more and more sharply peaked as $p \rightarrow 1$, the interpolation error is more and more concentrated in the vicinity of the origin, too. The Fourier or Chebyshev error is *highly nonuniform* when $f(x)$ has poles or branch points near the interpolation interval.

3. THE NEGLECT-OF-TRIPLE ALIASING (NTA) APPROXIMATION

Theorem 4: *Neglect-of-Triple-Aliasing (NTA) Approximation.* Let ρ_j denote the coefficients of the "envelope" of the interpolation error. Let α_j be the coefficients of the function $f(x)$ that is being interpolated. Then the approximation

$$\rho_j = 2\alpha_{j+N} + O(\alpha_{j+3N}), \qquad j = 0, 1,... \quad [\text{“NTA” approximation}] \qquad (3.1)$$

is accurate to within the absolute error indicated.

This approximation and theorem apply to an antisymmetric function without modification except for the purely notational substitutions $\alpha_j \rightarrow \beta_j$ and $\rho_j \rightarrow \sigma_j$.

If $f(x)$ is periodic and has no singularities for real x (Fourier case), or has no singularities on $x \in [-1, 1]$ (Chebyshev applications), then the error in (3.1) is roughly the *square* of the error in the interpolant itself, i.e., $O(E_I^2)$.

Proof. The proof is accomplished by truncation of the infinite series (2.13) after the leading term. □

This approximation is dubbed "Neglect-of-Triple-Aliasing" because all the error terms have degrees greater than $3N$. The singly aliased terms are those with $j \in [N, 2N-1]$, while the "doubly aliased" have $j \in [2N, 3N-1]$. Table I illustrates the accuracy of the NTA.

To understand the smallness of the error in this approximation, we need to review the qualitative theory of spectral expansions.

If the convergence of the series for a function $f(x)$ is limited by complex singularities (Fourier expansions) or singularities outside the real interval $x \in [-1, 1]$ [applications of (3.1) to Chebyshev series], then the series will convergence *geometrically*, that is,

$$|a_j| \leqslant [\ \ ]\, p^j \tag{3.2}$$

where p is a constant smaller than than unity and j is sufficiently large. The empty brackets stand for a factor that varies algebraically with j—much more slowly than p^j, which is an *exponential* function of j. {Note that $p^j = \exp[j \log(p)]$}. This algebraic factor depends upon the *type* of the singularity (simple pole, logarithm, cube root, etc.). In contrast, the constant p depends solely upon the *location* of the singularity; in the Fourier case, p is the logarithm of the imaginary part of the location of the singularity that is nearest the real x axis (Boyd, 1989a).

These elementary facts have two important implications. The first is that the model functions $\lambda(x; p)$ and $\mu(x; p)$ are *representative* of a very

Table I. Illustration of the Error in the Neglect-of-Triple-Aliasing (NTA) Approximation for $\lambda(x; p)$ for Various p and a Fixed Number of Iterpolation Points[a]

p	$\frac{\text{Max}(E_I)}{\text{Max}(\lambda)}$	$\frac{\text{Max(NTA)}}{\text{Max}(E_I)}$	$\text{Max}(E_I)$	Max(NTA)
0.6	7.3E−5	1.3E−9	2.9E−4	3.8E−13
0.65	3.6E−4	3.3E−8	1.7E−3	5.6E−11
0.7	1.6E−3	6.4E−7	9.0E−3	5.8E−9
0.75	6.3E−3	1.0E−5	0.044	4.5E−7
0.8	0.023	1.3E−4	0.208	2.8E−5
0.85	0.077	1.5E−3	0.95	1.4E−3
0.9	0.240	0.015	4.56	0.067
0.95	0.635	0.129	24.8	3.18

[a] The second column lists the maximum of the interpolation error E_I divided by the maximum of $\lambda(x; p)$ for a given p. The third column is the relative error in the NTA approximation: the maximum error in this approximation divided by the maximum in the interpolation error. Columns four and five give the absolute errors in the 20-point cosine interpolation and in the NTA approximation.

wide class of functions. We shall develop this theme in later sections by looking at many examples to show that their interpolation and truncation errors closely resemble those for $\lambda(x; p)$ and $\mu(x; p)$.

The second implication is the final proposition of the theorem: that the error is roughly the square of that of the interpolant. If the coefficients are decreasing roughly as p^j, then the interpolation and truncation errors will both be $O(p^N)$ as shown explicitly for $\lambda(x; p)$ and $\mu(x; p)$ in Theorem 3. Then, the errors in the NTA approximation will be $O(p^{3N})$ or smaller. For example, if N is large enough so that the interpolant agrees with $f(x)$ to three decimal places, the spectral coefficients in the error envelope will be misrepresented by the NTA approximation only in the ninth decimal place, a relative error of $O(10^{-6})$. For $\lambda(x; p)$ and $\mu(x; p)$, the NTA approximation is equivalent to replacing the factors of $1/(1+p^{2N})$ by 1 in Theorem 3.

Since our principal concern is to minimize E_I and E_T, and the error-of-the-error is secondary, the NTA approximation is blessed with more than adequate accuracy. Despite its simplicity, however, it is not simple enough. In the following sections, we develop other, simpler but less accurate approximations, which shall give deeper insights into the error in Fourier and Chebyshev interpolation.

4. NEGLECT-OF-DOUBLE-ALIASING: THE COEFFICIENT PAIRING OF THE SPECTRAL COEFFICIENTS OF THE INTERPOLATION ERROR

The following two theorems are a little out of place since they give *exact* results, and all the other exact formulas are collected in Sec. 2. We have postponed Theorems 5 and 6 until now because they are essential lemmas in the proof of the "coefficient-pairing" theorem below. The method of proof also helps to understand how the truncation and interpolation errors are related, which is a major theme of this section.

Theorem 5: *Aliasing Relations.* Let the points of the interpolation grid be given by

$$x_i = \pi(2i-1)/(2N), \qquad i = 1,..., N \tag{4.1}$$

We introduce the following notation: an equals sign with a "G" above it will denote an equality that is true only on the interpolation grid, and not for other values of x. Then we have

$$\cos(kx + 2Nmx) \overset{G}{=} (-1)^m \cos(kx), \qquad k, m \text{ integers}$$

$$(\text{"Aliasing Relation"}) \qquad (4.2)$$

$$\sin(kx + 2Nmx) \overset{G}{=} (-1)^m \sin(kx), \qquad k, m \text{ integers}$$

Table II illustrates the aliasing relations for $N=5$.

Proof. The first step is to observe that by an elementary trigonometric identity

$$\cos[(k+2Nm)x] = \cos(kx)\cos(2Nmx) + \sin(kx)\sin(2Nmx) \qquad (4.3)$$

However, on the interpolation grid (4.1),

$$\cos(2Nx_i) = (-1)^m \quad \text{and} \quad \sin(2Nmx_i) = 0, \qquad i = 1,\ldots,N \qquad (4.4)$$

Subsituting (4.4) into (4.3) gives the first part of (4.2). The proof of the second part is similar. □

Theorem 6: *Exact Low Degree Fourier Coefficients for the Interpolation Error.* Let $f(x)$ be a symmetric function whose exact Fourier coefficients are $\{\alpha_j\}$. Let $g(x)$ be a function such that $g(x) = -g(-x)$ for all x and let its Fourier sine coefficient be $\{\beta_j\}$. The low-degree coefficients of the error in interpolating these functions at the roots of $\cos(Nx)$ are

Table II.

(a) Aliasing for Cosine Interpolation at the Roots of Cos(Nx) for $N=5$[a]

1	$\cos(x)$	$\cos(2x)$	$\cos(3x)$	$\cos(4x)$	0	Unaliased
—	$-\cos(9x)$	$-\cos(8x)$	$-\cos(7x)$	$-\cos(6x)$	$-\cos(5x)$	Singly aliased
$-\cos(10x)$	$-\cos(11x)$	$-\cos(12x)$	$-\cos(13x)$	$-\cos(14x)$	—	Doubly aliased
—	$\cos(19x)$	$\cos(18x)$	$\cos(17x)$	$\cos(16x)$	$\cos(15x)$	Triply aliased
$\cos(20x)$	$\cos(21x)$	$\cos(22x)$	$\cos(23x)$	$\cos(24x)$	—	
—	$-\cos(29x)$	$-\cos(28x)$	$-\cos(27x)$	$-\cos(26x)$	$-\cos(25x)$	
$-\cos(30x)$	$-\cos(31x)$	$-\cos(32x)$	$-\cos(33x)$	$-\cos(34x)$	—	
—	$\cos(39x)$	$\cos(38x)$	$\cos(37x)$	$\cos(36x)$	$\cos(35x)$	

(b) Aliasing for Sine Interpolation at the Roots of cos(Nx) for $N=5$[a]

0	$\sin(x)$	$\sin(2x)$	$\sin(3x)$	$\sin(4x)$	$\sin(5x)$	Unaliased
—	$\sin(9x)$	$\sin(8x)$	$\sin(7x)$	$\sin(6x)$	—	Singly aliased
$-\sin(10x)$	$-\sin(11x)$	$-\sin(12x)$	$-\sin(13x)$	$-\sin(14x)$	—	Doubly aliased
—	$-\sin(19x)$	$-\sin(18x)$	$-\sin(17x)$	$-\sin(16x)$	$-\sin(15x)$	Triply aliased
$\sin(20x)$	$\sin(21x)$	$\sin(22x)$	$\sin(23x)$	$\sin(24x)$	—	
—	$\sin(29x)$	$\sin(28x)$	$\sin(27x)$	$\sin(26x)$	$\sin(25x)$	
$-\sin(30x)$	$-\sin(31x)$	$-\sin(32x)$	$-\sin(33x)$	$-\sin(34x)$	—	
—	$-\sin(39x)$	$-\sin(38x)$	$-\sin(37x)$	$-\sin(36x)$	$-\sin(35x)$	

[a] All expressions in a given column are *identical* when evaluated at the roots of $\cos(5x)$.

$$e_k = \sum_{m=1}^{\infty} (-1)^{m+1}(\alpha_{2Nm-k} + \alpha_{2Nm+k}), \qquad k=0,..., N-1 \quad \text{(cosine)} \qquad (4.5)$$

$$e_k = \sum_{m=1}^{\infty} (-1)^{m}(\beta_{2Nm-k} - \beta_{2Nm+k}), \qquad k=1,..., N-1 \quad \text{(sine)}$$
$$e_N = \sum_{m=1}^{\infty} (-1)^{m+1}\beta_{2Nm+N} \quad \text{(sine)} \qquad (4.6)$$

The low-degree coefficients for $E_T(x; N)$ are identically zero.

This theorem is complementary to Theorem 1 (Sec. 2), which showed that all the higher Fourier coefficients for both the interpolation and truncation errors are identical with the corresponding coefficients of the functions being interpolated.

Proof. The interpolation error can be expressed in terms of the truncation error as

$$E_I(x; N) = E_T(x; N) - I_N E_T(x; N) \qquad (4.7)$$

To justify (4.7), note that the expression on the right-hand side of (4.7) meets two requirements. First, the difference between a function and its interpolant always vanishes at the interpolation points, as E_I must. Second, the high-degree coefficients of E_I and E_T match, as required by Theorem 1, because only a low-degree polynomial, $I_N E_T$, is subtracted from E_T.

The prescription for constructing the interpolating polynomial $I_N E_T$ in (4.7) is simple. For any function $f(x)$ we wish to interpolate, modify its Fourier series by replacing the trigonometric function of each term by its alias via (4.4), then add all the aliases together to obtain $I_N f$. (For the Fourier terms of low degree, i.e., $m \leqslant N$, the alias is the trigonometric function itself so that the alias replacement does nothing.) Each trigonometric term has an infinite number of aliases, but because of the factor of $2N$ in the argument of $\cos(kx + 2mNx)$ and the restriction of m to positive and negative integers (and zero), each cosine function has a *unique* alias *within* the set $\{1, \cos(x),..., \cos[(N-1)x]\}$ and similarly each sine has a unique alias within the span of $\{\sin(x),..., \sin(Nx)\}$. The interpolant is the sum of these low-degree aliases. Applying this interpolation procedure to (4.7) after obtaining the Fourier series of $E_T(x; N)$ from Theorem 1 then proves Theorem 6. □

In the next theorem, we turn from exact formulas to approximations.

Theorem 7: *Neglect-of-Double-Aliasing (Coefficient-Pairing) Approximation.* If the spectral coefficients of a symmetric function $f(x)$ are denoted $\{\alpha_j\}$ while those of the cosine interpolation error $E_I(x; N)$ are denoted $\{e_j\}$, then

$$e_0 \approx 2e_{2N}\{1 + \alpha_{4N}/\alpha_{2N} - \cdots\} \qquad \text{(cosine)} \tag{4.8a}$$

$$e_j \approx e_{2N-j}\{1 + \alpha_{2N+j}/\alpha_{2N-j} - \cdots\}, \qquad j = 1,..., N-1 \tag{4.8b}$$

which is equivalent to

$$E_I \approx \alpha_N \cos(Nx) + \sum_{j=1}^{N} \alpha_{N+j}\{\cos[(N+j)x] + \cos[(N-j)x]\} + O(\alpha_{2N+1}) \tag{4.9}$$

The special case $e_0 = 2\alpha_{2N}$ is actually not an exception to the general rule, (4.8b). By convention this term is halved in evaluating the sum as explicit in (2.8), so the net contribution of e_0 to the interpolation error does indeed approximately equal that of e_{2N}.

Similarly, for sine interpolation

$$e_j \approx -e_{2N-j}, \qquad j = 1,..., N-1 \tag{4.10}$$

$$\text{(sine)}$$

$$E_I(x; N) \approx \sum_{k=1}^{N-1} \beta_{N+k}\{\sin[(N+k)x] - \sin[(N-k)x] \tag{4.11}$$

Proof. Neglect of all terms of degree greater than $2N$ in Theorem 6 (which gives the coefficients for $k < N$) combined with Theorem 1 (which specifies the coefficients for $k > N$). We refer to this as "Neglect-of-Double-Aliasing" because the only surviving terms in (4.8)–(4.11) are the unaliased and singly aliased terms, i.e., those of degree less than N and less than $2N$, respectively.

An equivalent and simpler argument is to truncate the series for the *envelope* to its first N terms. Invoking Theorem 4, (3.1), to make the substitution $\rho_j \to 2\alpha_{N+j}$ and then applying the identity

$$2\cos(Nx)\cos(jx) = \cos[(N+j)x] + \cos[(N-j)x], \qquad j \leqslant N \tag{4.12}$$

gives (4.9). For sine interpolation, we similarly expand $E_I(x; N) = \cos(Nx)\,\sigma(x; N)$ using the identity

$$2\cos(Nx)\sin(jx) = \sin[(N+j)x] - \sin[(N-j)x], \qquad j \leqslant N-1 \tag{4.13}$$

For a geometrically converging series, the error in the interpolation error is $O(p^N)$ while the *relative* error in Theorem 7 is of the same order. Thus, this theorem is a very safe and accurate approximation.

Collectively, Theorems 6 and 7 emphasize the notion that in some sense, the interpolation error is double the truncation error. "In some sense" means that each high-degree Fourier coefficient ($N < m \leqslant 2N$) of the function being interpolated, $f(x)$, appears *once* in the truncation error $E_T(x; N)$ while it appears *twice* in the interpolation error—once as itself and once as its alias. Note that the first term in the square brackets in (4.9) appears in both E_T and E_I, while the second, also weighted by α_{N+j}, appears only in the interpolation error.

Although (4.9) is only approximate, it does allow a very simple and visual interpretation of the difference between the interpolation and truncation errors. Figure 5 compares the first $2N$ coefficients of both errors.

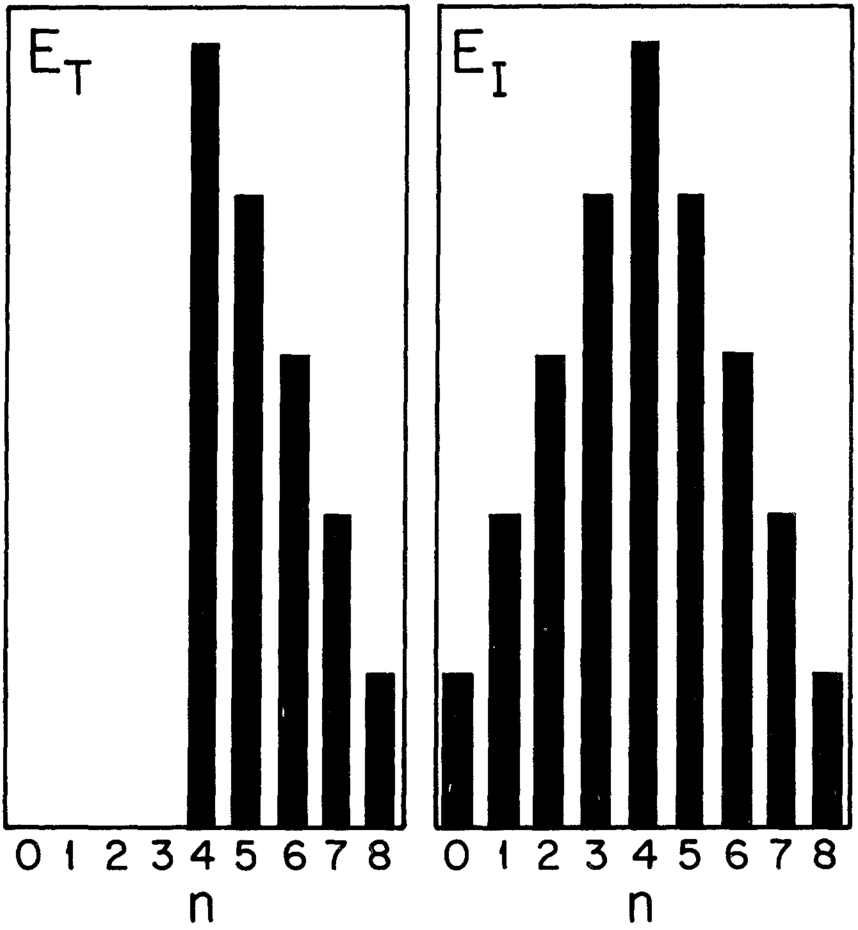

Fig. 5. Schematic of $\ln|a_n|$ versus n for the Chebyshev (or Fourier cosine) coefficients of the truncation error (left panel) and interpolation error (right side) for a typical analytic function $f(x)$. (The logarithms have been scaled by subtracting $\ln|a_9|$ so that the smallest coefficient shown is represented by the shortest bar.)

On the semilogarithmic graph, the bars that represent the coefficients of the interpolation error can be bounded by an isosceles triangle. The coefficients of the truncation error can be bounded by the right half of this same triangle. Each coefficient in the interpolation error is paired with its alias [except for that of $\cos(Nx)$].

Because of the interference effects between different Fourier terms at different values of x, the *pointwise* values of E_I and E_T need not be in simple constant ratio, as illustrated by Fig. 4, but the *maximum* of $|E_I|$ is no more than double the maximum absolute value of the truncation error.

Table III is a numerical illustration of "coefficient pairing." For a geometrically converging series in which each term is proportional to p^j, Theorem 7 can be more precisely expressed as

$$e_j \approx e_{2N-j}[1 + p^{2j} + O(p^{2N})], \qquad j = 1, ..., N-1 \tag{4.14}$$

as illustrated in this table.

Table III. An Illustration of Coefficient Pairing in the Interpolation Error[a]

j	e_j	e_{2N-j}	$2N-j$	$e_j - e_{2N-j}$	α_{2N+j}
	$e_{20} = 0.00159585$				
19	0.0011170925	0.0011170917	21	0.0000000007	0.0000000015
18	0.0007819658	0.0007819642	22	0.0000000016	0.0000000021
17	0.0005473776	0.0005473749	23	0.0000000026	0.0000000030
16	0.0003831665	0.0003831625	24	0.0000000040	0.0000000042
15	0.0002682196	0.0002682137	25	0.0000000059	0.0000000060
14	0.0001877581	0.0001877497	26	0.0000000085	0.0000000086
13	0.0001314370	0.0001314247	27	0.0000000123	0.0000000123
12	0.0000920149	0.0000919973	28	0.0000000176	0.0000000176
11	0.0000644233	0.0000643981	29	0.0000000251	0.0000000252
10	0.0000451146	0.0000450787	30	0.0000000359	0.0000000360
9	0.0000316064	0.0000315551	31	0.0000000514	0.0000000514
8	0.0000221619	0.0000220886	32	0.0000000734	0.0000000734
7	0.0000155668	0.0000154620	33	0.0000001049	0.0000001049
6	0.0000109732	0.0000108234	34	0.0000001498	0.0000001498
5	0.0000077904	0.0000075764	35	0.0000002140	0.0000002140
4	0.0000056092	0.0000053035	36	0.0000003057	0.0000003057
3	0.0000041492	0.0000037124	37	0.0000004368	0.0000004368
2	0.0000032226	0.0000025987	38	0.0000006239	0.0000006239
1	0.0000027104	0.0000018191	39	0.0000008914	0.0000008914
0	0.0000025467	0.0000012734	40	0.0000012734	0.0000012734

[a] The second and third columns are the coefficients of the error in the 20-point cosine interpolation of $\lambda(x; p = 0.7)$. (The unpaired coefficient of $\cos(Nx)$ is listed on top of these two columns.) The first and fourth columns give the degrees of the coefficients. The fifth column is the difference between the paired coefficients. The sixth column is the prediction for this difference from Theorem VI:

$$e_j \approx e_{2N-j} + \alpha_{2N+j} - \alpha_{4N-j} + \cdots.$$

5. TAYLOR SERIES FOR THE LOGARITHM OF THE COEFFICIENTS: THE EFFECTIVE LORENTZIAN APPROXIMATION

The "effective Lorentzian" approximation has an error which is usually $O(1/N)$. This is much poorer than the earlier neglect-of-higher-aliasing approximations, which have an error which is an exponential function of N. Nevertheless, the effective Lorentzian is the most useful of the approximations we will discuss. Its success is an illustration of C. S. Yih's Law of Inverse Usefulness: The more useful an approximation, the larger its error.

Theorem 8: *Effective Lorentzian Approximation.* If the logarithm of the Fourier coefficient is sufficiently smooth, then by Taylor expansion of the logarithm (and the Neglect-of-Triple Aliasing approximation), we have

$$\rho_j \approx 2\alpha_N e^{j\alpha'(N)/\alpha(N)}[1+O(j^2)] \tag{5.1}$$

$$\approx 2\alpha_N (p_{\text{eff}})^j \tag{5.2}$$

where the $\{\alpha_j\}$ are the coefficients of $f()x)$, the function being interpolated, the $\{\rho_j\}$ are the coefficients of the envelope of the interpolation error, the prime denotes differentiation, and

$$p_{\text{eff}} \equiv e^{\alpha'(N)/\alpha(N)} \tag{5.3}$$

The reason for defining p_{eff} is that (5.2) asserts that the coefficients of the envelope of the interpolation error are approximately those of the Lorentzian function

$$\rho(x; N, f) \approx \alpha_N \lambda(x; p_{\text{eff}}) \tag{5.4}$$

The theorem applies to sine interpolation without modification except for the replacement of the symmetric Lorentzian λ by the antisymmetric serpentine $\mu(x; p)$.

Proof. Taylor expansion gives

$$\log(\alpha_{N+j}) = \log(\alpha_N) + j(d\alpha/dj)/\alpha(N) + O(j^2), \qquad j \ll N \tag{5.5}$$

Substituting this into the identity

$$\alpha_{N+j} = e^{\log(\alpha_{N+j})} \approx e^{\log(\alpha_N)} e^{j\alpha'(N)/\alpha(N)} + O(j^2) \tag{5.6}$$

and inserting (5.6) into Theorem 4 gives (5.1).

The "effective Lorentzian" approximation has severe limitations. In particular, it cannot be applied if the Fourier coefficients are rapidly *oscillating* with respect to *degree.* (Note that a generalization of Theorem 8 that does apply to oscillating Fourier coefficients is defined in the next section.) Nevertheless, the effective Lorentzian approximation does encourage us to generalize our range of model functions.

Define

$$\lambda_k(x;p) \equiv 2\sum_{j=1}^{\infty} j^k p^j \cos(jx) \tag{5.7a}$$

$$\mu_k(x;p) \equiv 2\sum_{j=1}^{\infty} j^k p^j \sin(jx) \tag{5.7b}$$

Since

$$\alpha'(N)/\alpha(N) = -\log(p) + k/N \tag{5.8a}$$

$$\approx -\log(p) + O(1/N) \tag{5.8b}$$

$p_{\text{eff}} = p$ for these functions and the effective Lorentzian approximation is

$$E_I(x;p,\lambda_k) \approx 2N^k p^N \cos(Nx)\,\lambda(x;p) \tag{5.9a}$$

$$E_I(x;p,\ \mu_k) \approx 2N^k p^N \cos(Nx)\,\mu(x;p) \tag{5.9b}$$

For positive integer k, the functions λ_k and μ_k are the kth derivatives of either $\lambda(x;p)$ or $\mu(x;p)$. Thus, they have poles of $(k+1)$st order in the complex plane, that is,

$$\lambda_k(x;p) \sim (\text{const})/[x - i\log(p)]^{k+1} \qquad , |x - i\log(p)| \ll 1 \tag{5.10}$$

Equation (5.9) shows that the *order* of the poles affects the interpolation error through a *power* of N. The exponential factor of N in (5.9), however, is $p^N = \exp[-N\log(p)]$, which depends only upon the *location* of the poles. Similarly, the shape of the error envelope is not affected by the *order* of the singularity, only by its *location.*

In Sec. 7, we shall return to these generalized model functions with graphs and a theorem expressing the exact interpolation error.

As noted earlier, the effective Lorentzian approximation is not applicable to all classes of functions; it fails whenever the series coefficients oscillate rapidly (with an exception explained in the next section). This approximation does apply whenever the Fourier or Chebyshev coefficients are of the form

$$\alpha_n \sim dn^k e^{-sn^r} \tag{5.11}$$

for some constants d, k, s, and r. When $r=1$, these coefficients converge geometrically and are those of the model function λ_k.

In Sec. 7, we shall describe important model functions for other values of the "exponential index of convergence" r.

One limitation of all the cases described so far is that the poles and branch points are limited to the imaginary x axis so that the series coefficients decay monotonically. In the next section, we show that shifting the singularities parallel to the real axis does not qualitatively alter the interpolation error.

6. PHASE SHIFTING AND THE INTERPOLATION ERROR

Theorem 9: *Interpolation Error for the Phase-Shifted Lorentzians.* Let ρ_λ and ρ_μ denote the envelopes of the interpolation error for the phase-shifted Lorentzian functions, $\lambda(x+\phi; p)$ and $\mu(x+\phi; p)$, respectively, where the phase shift ϕ is real. Then these error envelopes in the "Neglect-of-Triple-Aliasing" approximation are given by

$$\left|\begin{matrix} \rho_\lambda(x; p, N) \\ \rho_\mu(x; p, N) \end{matrix}\right| = 2p^N \left|\begin{matrix} \cos(N\phi) & -\sin(N\phi) \\ \sin(N\phi) & \cos(N\phi) \end{matrix}\right| \left|\begin{matrix} \lambda(x+\phi; p) \\ \mu(x+\phi; p) \end{matrix}\right| \tag{6.1}$$

within a relative error of $O(p^{2N})$.

The 2×2 matrix in (6.1) is the matrix that describes the transformation of Cartesian coordinates under rotation through an angle by $(-N\phi)$ radians.

Proof. By using the definitions of λ and μ and elementary trigonometric identities, we obtain

$$\begin{aligned} \lambda(x+\phi; p, N) = 1 + 2 \sum_{j=1}^{\infty} p^j \cos(j\phi) \cos(jx) \\ -2 \sum_{j=1}^{\infty} p^j \sin(j\phi) \sin(jx) \end{aligned} \tag{6.2}$$

$$\begin{aligned} \mu(x+\phi; p, N) = 2 \sum_{j=1}^{\infty} p^j \sin(j\phi) \cos(jx) \\ +2 \sum_{j=1}^{\infty} p^j \cos(j\phi) \sin(jx) \end{aligned} \tag{6.3}$$

Note that these coefficients *oscillate* with degree j; Theorem 9 extends the effective Lorentzian approximation to (some) functions with oscillating Fourier coefficients.

The triple aliasing approximation implies that the interpolation error envelope is

$$\begin{aligned} \rho(x; N, \lambda(x+\phi; p)) \approx {} & 4 \sum_{j=1}^{\infty} p^{N+j} \cos[(N+j)\phi] \cos(jx) \\ & -4 \sum_{j=1}^{\infty} p^{N+j} \sin[(N+j)\phi] \sin(jx) \qquad (6.4) \\ = {} & 2p^N \left\{ \cos(N\phi) \left[1 + 2 \sum_{j=1}^{\infty} p^j \cos(j\phi) \cos(jx) \right] \right. \\ & - \sin(N\phi) 2 \sum_{j=1}^{\infty} p^j \sin(j\phi) \cos(jx) \\ & - \sin(N\phi) 2 \sum_{j=1}^{\infty} p^j \cos(j\phi) \sin(jx) \\ & \left. - \cos(N\phi) 2 \sum_{j=1}^{\infty} p^j \sin(j\phi) \sin(jx) \right\} \qquad (6.5) \end{aligned}$$

Collecting terms and applying (6.2) and (6.3) gives the first line of the theorem. The error estimate for $\mu(x+\phi; p)$ is similar.

If the shift ϕ is a multiple of the grid spacing, $h = \pi/N$, then nothing is altered except that $\lambda(x) \rightarrow \lambda(x+\phi)$ both for the function being approximated and for the envelope of the interpolation error. (This could have been predicted in advance because, for trigonometric interpolation, all grid points are equivalent.) If the shift is a half-integral multiple of h, then ρ is proportional to $\mu(x+\phi; p)$—but $\mu(x+\phi; p)$ is also sharply concentreated near $x = -\phi$. (Intermediate ϕ give an envelope that is a hybrid of these two cases.)

It follows that the degree of nonuniformity in x is only mildly affected by the shift. The shape of the error envelope is controlled primarily by the *imaginary part* of the position of those poles and branch points that are closest to the real x axis. The *real part* of the location of the singularity merely *shifts* the *maximum* of the error envelope without much affecting its shape or magnitude. "Without much" means that a phase shift in $f(x)$ can convert the shape of the envelope from a Lorentzian to a serpentine or vice versa, but the envelope is always a maximum near the pole or the branch point.

It is intriguing that the matrix in (6.1) has the form of a rotation matrix. The interpolation error for the Lorentzian function is always proportional to a linear combination of these two functions. The phase shift ϕ merely alters the proportion. Because the determinant of the rotation matrix is unity, i.e., the matrix multiplication is a pure rotation without amplification, the norm of the interpolation error is not significantly changed by a phase shift.

7. EXAMPLES OF THE "EFFECTIVE LORENTZIAN" APPROXIMATION

7.1. "Supergeometric Convergence": Riemann Theta Function

The Lorentzian functions have "geometric" convergence, i.e., coefficients proportional to p^j for some p such that $|p| < 1$, because these functions have complex singularities that are not at infinity. Entire functions, which have no singularities except at infinity, have a "supergeometric" rate of convergence in the language of Boyd (1989). It is remarkable that the error in interpolating such functions is still approximately proportional to the Lorentzian or serpentine functions.

As an example, consider the theta function θ_3, which is important in theories of diffusion, elliptic functions, and solitary waves:

$$\theta_3(x/2; q) = 1 + 2 \sum_{n=1}^{\infty} q^{n^2} \cos(nx) \tag{7.1}$$

The rate of convergence is extraordinarily fast: the nth coefficient is proportional to the n^2-power of a constant q versus the nth power of a constant for a geometrically convergent Fourier series. Indeed, the theta series converges unusually fast even for an entire function: the usual situation is coefficients of $O[\exp(-qn \log n)]$ for entire functions (Boyd, 1989a).

Nevertheless, the effective Lorentzian approximation still applies, predicting the error envelope

$$\rho(x; q, N) \approx 2q^{N^2} \lambda(x; p_{\text{eff}}) \tag{7.2}$$

where

$$p_{\text{eff}} \equiv q^{2N} \tag{7.3}$$

Figure 6 compares the shape of the theta function with the exact error envelope and the approximate envelope given by (7.2). There are two striking conclusions.

The first is that the "effective Lorentzian" approximation is very accurate. The coefficients of the theta function decrease smoothly, monotonically, and very rapidly so that the Taylor expansion (5.8) is

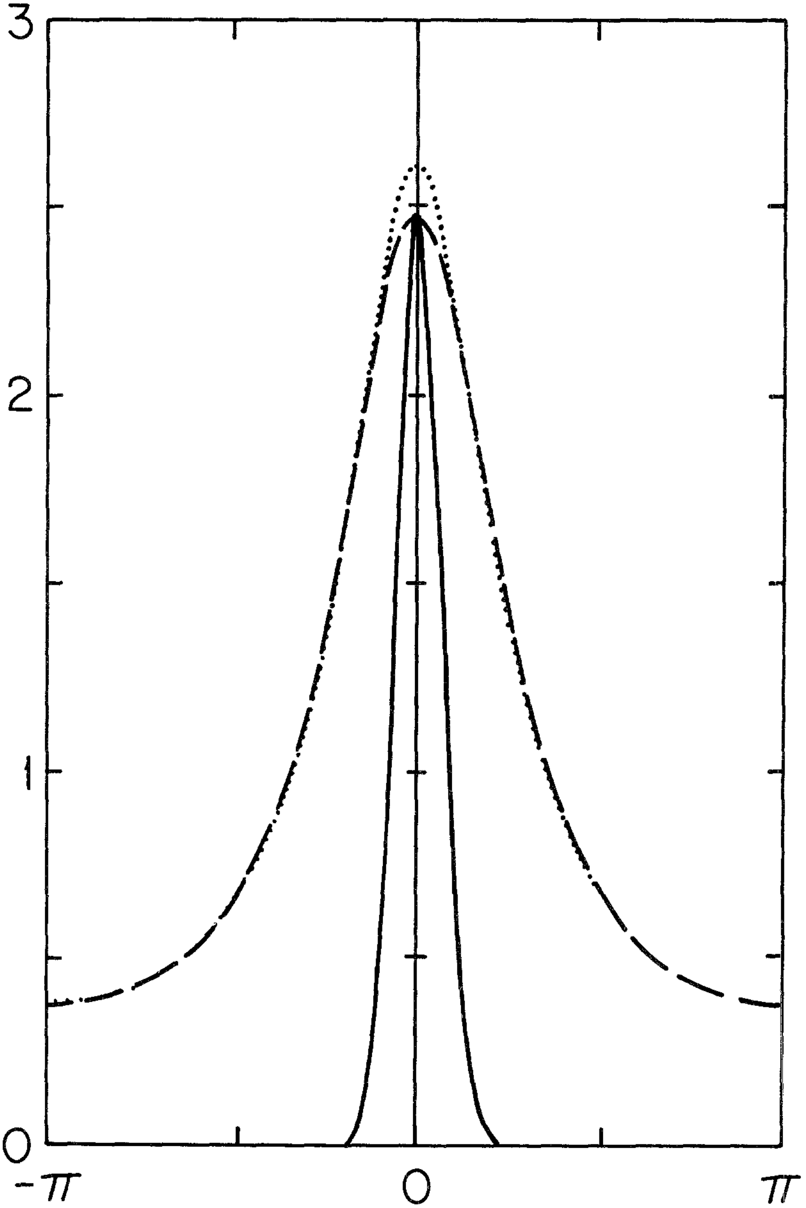

Fig. 6. Solid curve: $\theta_3(x/2)$ with the coefficients $a_n = 2q^{n^2}$ with $q = 0.98$. Dashed curve: exact envelope of the interpolation error. Dotted curve: approximation to the envelope which is proportional to $\lambda(x, p_{\text{eff}})$ where $p_{\text{eff}} = q^{2N}$. The two error curves have been divided by a_N while θ_3 has been divided by 5 so that one may compare shapes on the same curve. $N = 20$.

accurate until the coefficients of the theta function have become negligibly small.

The second interesting feature of the graph is that the theta function is much narrower than the Lorentzian. The error is peaked about the maximum of the function that is being interpolated, but the interpolation of the entire function has smoothed the error in the sense that the error is much more uniform in x than is the theta function itself.

We shall limit ourselves to a theta function that has its maximum at $x=0$ because as shown for Lorentzian functions in the previous section, shifting the function that is being interpolated shifts the error envelope without significantly altering its magnitude.

7.2. "Subgeometric Convergence": Infinitely Differentiable at the Branch Point

A function whose Fourier coefficients are decreasing as $O[\exp(-qj^r]$ for some constants q and r is said to have "subgeometric convergence" if $r<1$. (The most common case, $r=1$, is "geometric convergence," while $r>1$ is "supergeometric convergence.") The significance of subgeometric convergence is that the Fourier functions $\cos(jx)$ blow up proportional to $\cosh[j\,\mathrm{Im}(x)]$ away from the real x axis. When $r<1$, this fast growth of the basis functions with j for any x off the real axis cannot be overcome by the exponential decrease of the coefficients. The result is that the Fourier series *converges exponentially fast* for *real x*, but *diverges everywhere off the real x axis.*

Subgeometric convergence implies (Boyd, 1989a) that $f(x)$ must be nonanalytic for some real x: if the only singularities were off the real x axis with the closest ones at $\mathrm{Im}(x)=\pm a$, then the Fourier series would converge within the strip $|\mathrm{Im}(x)|<a$. However, if the Fourier coefficients decrease exponentially fast with n, even if subgeometrically, then the series—and $f(x)$—can be differentiated an arbitrary number of times for real x without series divergence or infinities. In mathematical jargon, a function with subgeometric convergence is "C^∞" (i.e., infinitely differentiable) but not "C^Ω" (analytic for all real x).

A typical example is defined by

$$\mathrm{SG}(x) \equiv 1 + 2\sum_{j=1}^{\infty} e^{-j^{2/3}} \cos(jx) \tag{7.4}$$

No simple closed form representation is known, but one can show that $\mathrm{SG}(x)$ is nonanalytic at the origin: Its power series about the origin diverges factorially.

In the "effective Lorentzian" approximation, the error is

$$E_I(x;\, N) \approx 2e^{-N^{2/3}} \cos(Nx)\, \lambda(x;\, p_{\text{eff}}) \tag{7.5}$$

$$p_{\text{eff}}(N) \equiv e^{-(2/3)N^{-1/3}} \tag{7.6}$$

Figure 7 shows that the periodic Lorentzian (7.5) approximates the interpolation error to within 4% of the maximum of the error.

It is striking that in contrast to the theta function, which was more

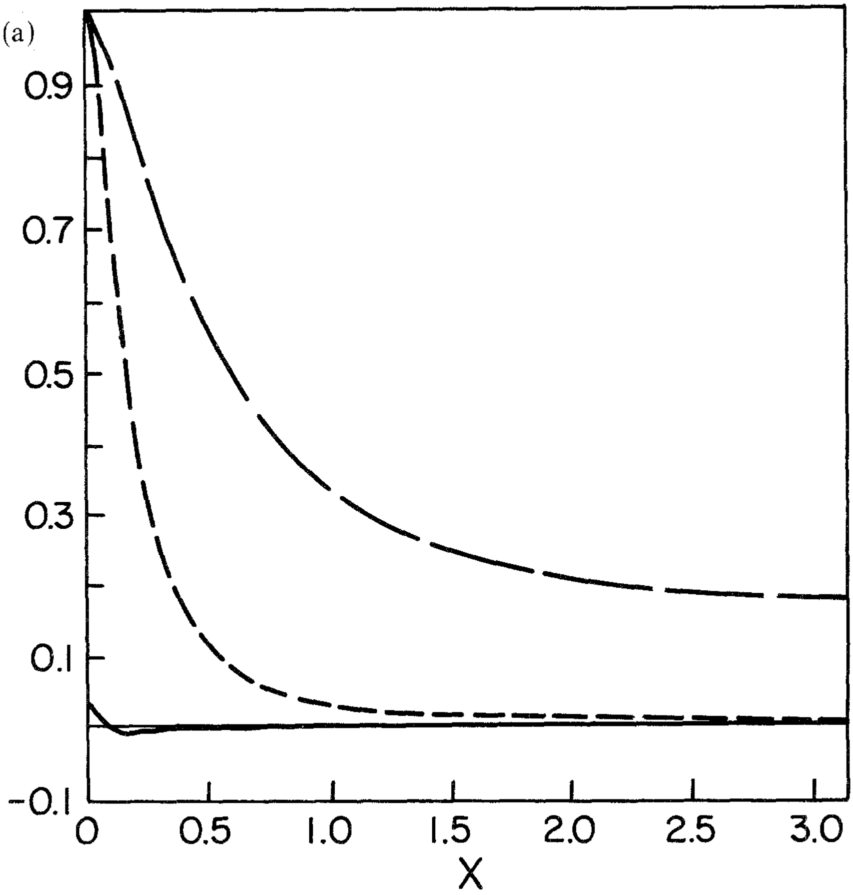

Fig. 7. (a) Long dashed curve: the function SG(x) whose coefficients are $a_n = 2\exp(-n^{2/3})$, divided by its maximum, SG(0) = 2.93. Short dashed curve: exact error envelope $\rho(x;\, N = 50)$, also scaled by its maximum value of 2.93E−5. The envelope as given by the effective Lorentzian approximation is not shown because it is indistinguishable from the exact envelope only in a very narrow zone around $x = 0$. Solid curve: the difference between the exact interpolation error envelope and the effective Lorentzian approximation to it. The same curve is repeated as Fig. 7b.

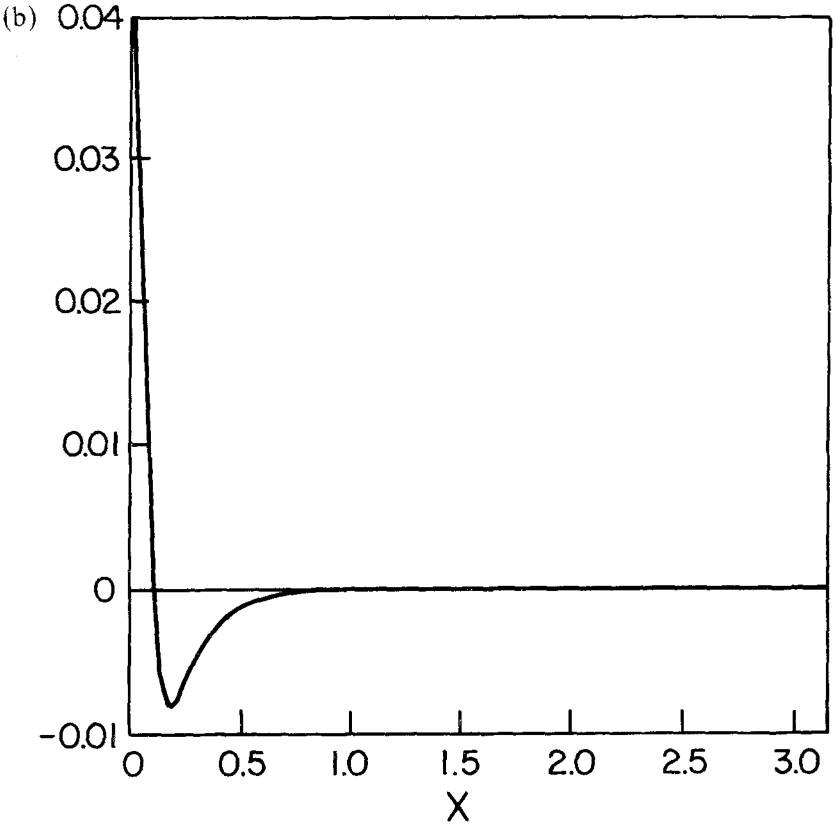

Fig. 7. Continued

sharply peaked than the interpolation error, the opposite is true here. The singularity at $x=0$ creates an interpolation error that is more concentrated at the origin than SG(x) itself. The parameter $p_{\text{eff}}(N) \to 1$ as $N \to \infty$, implying that the interpolation error becomes more and more sharply peaked as the number of grid points increases. This is hardly surprising: the singularity at $x=0$ is controlling the slow, subgeometric rate of convergence of the Fourier coefficients. This singularity also controls the rate of convergence of the interpolation error.

7.3. Geomerically Converging Functions: Poles and Branch Points

Figure 8 illustrates the exact and "effective Lorenztian" interpolation error envelopes for the function

$$\lambda_2 \equiv 2 \sum_{j=1}^{\infty} j^2 p^j \cos(jx) \tag{7.7}$$

which is a special case of the class of model functions defined by (5.7). The error in the "effective Lorentzian" approximation is

$$E_I(x; p, N) \approx 2N^2 p^N \cos(Nx)\, \lambda(x; p) \tag{7.8}$$

where, as for all geometrically converging series, $p_{\mathrm{eff}} = p$.

This model function has third-order poles at $x = \pm i \log(p)$ versus the first-order poles of $\lambda(x; p)$ at the same location, so $\lambda_2(x; p)$ is more sharply peaked about the origin than the Lorentzian. Figure 8 shows that the same is true of the interpolation error: (7.8) errs by predicting an error envelope which is a bit too small and too flat.

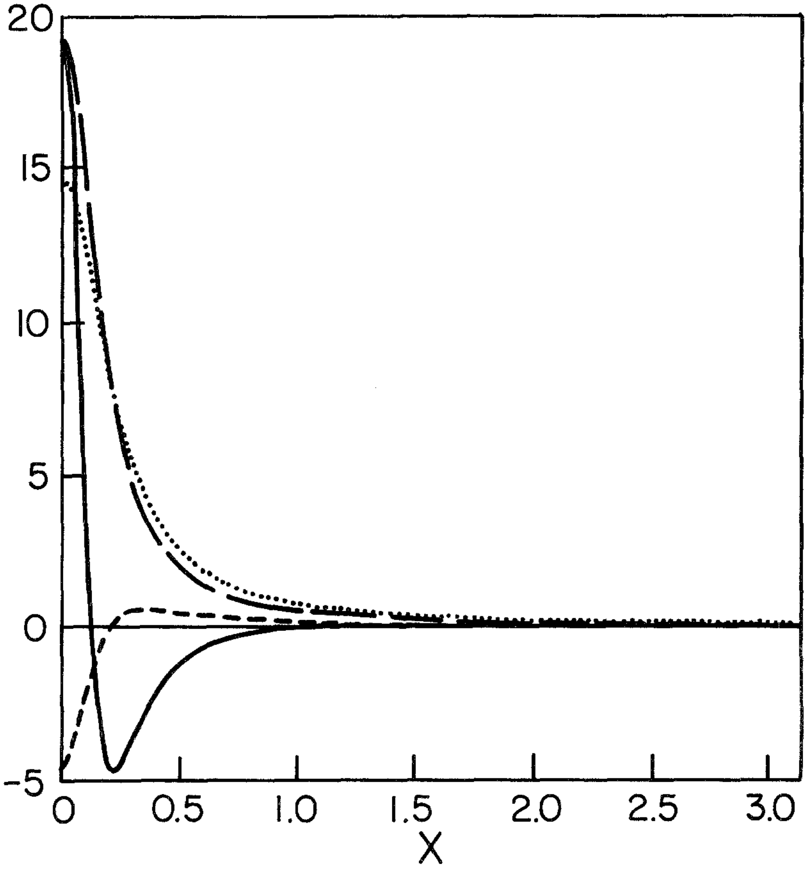

Fig. 8. Solid curve: $\lambda_2(x; p = 0.8)$ multiplied by 0.053 (so that its maximum is identical with the maximum in the interpolation error.) Long dashes: envelope of the exact interpolation error, $E_I(x; N)/\cos(Nx)$, for $N = 32$. Dotted curve: envelope of the interpolation error as predicted by the effective Lorentzian approximation, (7.8). Short dashes: the error in the effective Lorentzian approximation.

Although the "effective Lorentzian" approximation is not bad, we can eliminate most of the error-in-the-error by adding the leading term in the exact analytical formula for the error envelope which is given by the following:

Theorem 10: *Interpolation Error for Generalized Model Functions in the Neglect-of-Triple-Aliasing (NTA) Approximation.* For the model functions defined by

$$\lambda_k(x;\, p) \equiv 2 \sum_{j=1}^{\infty} j^k p^j \cos(jx) \tag{5.7a$'$}$$

$$\mu_k(x;\, p) \equiv 2 \sum_{j=1}^{\infty} j^k p^j \sin(jx) \tag{5.7b$'$}$$

the interpolation error in the NTA approximation (but without additional approximations) is given by

$$\lambda_k\!: \quad E_I(x;\, N) \approx 2N^k p^N \cos(Nx) \left\{ \lambda(x;\, p) + \sum_{m=1}^{\infty} [\Gamma(k+1)/\Gamma(m+1)\,\Gamma(k-m+1)](1/N)^m\, \lambda_m(x;\, p) \right\} \tag{7.9}$$

$$\mu_k\!: \quad E_I(x;\, p) \approx 2N^k p^{N+1} \cos(Nx) \left\{ \mu(x;\, p) + \sum_{m=1}^{\infty} [\Gamma(k+1)/\Gamma(m+1)\,\Gamma(k-m+1)](1/N)^m\, \mu_m(x;\, p) \right. \tag{7.10}$$

Proof. The proof is accomplished using the binomial theorem combined with the NTA approximation, Theorem 4. □

When k is a positive integer, the binomial series terminates after the kth term. Thus, in the Neglect-of-Triple-Aliasing approximation, the interpolation error for $\lambda_2(x;\, p)$ is

$$E_I(x;\, p) = 2N^2 p^N \cos(Nx)[\lambda(x;\, p) + (2/N)\, \lambda_1(x;\, p) + (1/N^2)\, \lambda_2(x;\, p)] \tag{7.11}$$

Figure 9 shows that most of the error in the "effective Lorentzian" approximation is removed by adding the correction proportional to $\lambda_1(x;\, p)$.

Our second example of geometric convergence is

$$A_{-3/2} \equiv \sum_{m=-\infty}^{\infty} \exp\{-\beta[a^2+(x-2\pi m)^2]^{1/2}\} \tag{7.12}$$

$$= (a/\pi)\, K_1(\beta a) + \sum_{j=1}^{\infty} \alpha_j \cos(jx) \tag{7.13}$$

where

$$\alpha_j \equiv 2\beta(a/\pi)\, K_1(a[\beta^2+j^2]^{1/2})/(\beta^2+j^2)^{1/2} \tag{7.14}$$

$$\sim 2p^j/j^{3/2} \qquad \text{for } j \gg 1, \qquad \beta=(2\pi/a)^{1/2} \tag{7.15}$$

where $a = \log(1/p)$. This function is *not* one of the models described by Theorem 10. However, when β is chosen as in (7.15), the Fourier

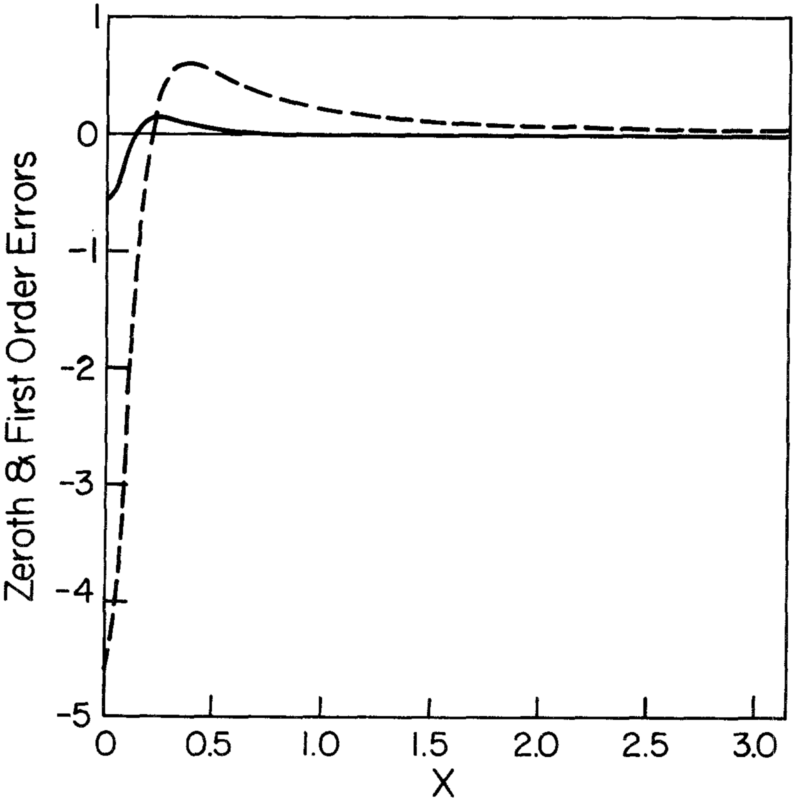

Fig. 9. Dashed curve: minus the error of the effective Lorentzian approximation to the interpolation error for $\lambda_2(x;\ p=0.8)$ (shown as the short-dashed curve in Fig. 8). Solid curve: the error-in-the-interpolation-error when E_I is approximated by a weighted sum of $\lambda(x;\ p)$ and $\lambda_1(x;\ p)$, (7.11). This correction term reduces the maximum error by a factor of 8.

coefficients of $\Lambda_{-3/2}(x; p)$ are *asymptotic* to those of $\lambda_{-3/2}(x; p)$. The gravest singularities for both functions are square root branch points at $x = \pm ia$.

Because only the high-degree Fourier coefficients enter the error estimates, we can still apply the effective Lorentzian approximation to $\Lambda_{-3/2}(x; p)$ as illustrated in Fig. 10.

7.4. Fractals and Functions with Natural Boundaries

The examples above are unrepresentative in one respect: each has only a single pair of singularities with complex conjugate locations. Fortunately,

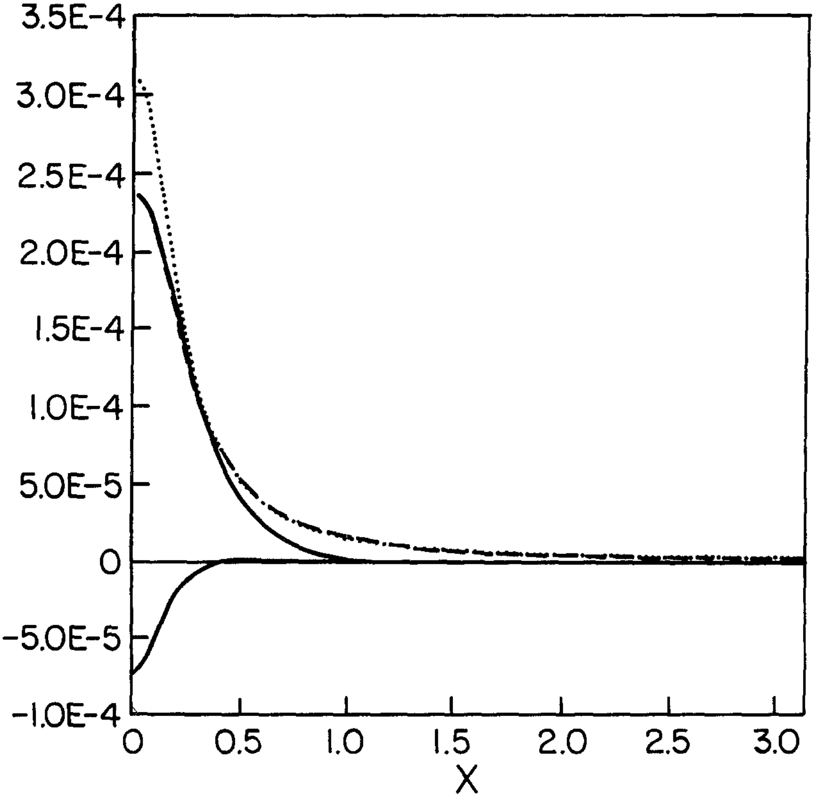

Fig. 10. Solid curve: $0.000771\,\Lambda_{-3/2}(x;\ p=0.8)$. (The numerical factor is chosen so that the maximum of the rescaled function is equal to the maximum in the interpolation error.) Dashed curve: exact envelope of the interpolation error for $N=27$ points. Dotted curve: effective Lorentzian approximation to the envelope of the interpolation error. Negative solid curve: error of the effective Lorentzian approximation; the maximum error is about 31% of the maximum in E_I.

this restriction can be removed by simply superimposing models with branch points at different places.

However, when the function has a continuous distribution of singularities, forming a so-called "natural barrier" in the complex plane beyond which the function cannot be analytically continued, then the function's spectral series is qualitatively different from the earlier models.

First, consider the function

$$\omega(x;p) \equiv \sum_{j=0}^{\infty} (1/3)^j \operatorname{sech}(3^j a) \cos(3^j x) \tag{7.16}$$

Since the coefficient of $\cos(kx)$ is bounded by p^k where $p = \exp(a)$, this series has geometric convergence, and can be bounded term-by-term by the terms of $\lambda(x;p)$. This in turn implies that the Fourier series converges exponentially fast not only for real x, but also for complex x such that $|\operatorname{Im}(x)| < a$.

Along the lines that bound the strip of convergence, $\omega(x;p)$ is not singular at a single isolated pole or branch point. Instead,

$$\operatorname{Re}\left\{\omega(x+ia;p) = \sum_{j=0}^{\infty} (1/3)^j \cos(3^j x)\right\} \equiv \Xi(x) \tag{7.17}$$

where we have used $\operatorname{Re}[\cos(x+ika)] = \cos(x)\cosh(ka)$. Equation (7.17) is Weierstrass's famous example of a function $\Xi(x)$, which is everywhere continuous but *nowhere differentiable* (Voss, 1988). Thus, $\omega(x;p)$ is singular *everywhere* along the lines $\operatorname{Im}(x) = \pm a$. These lines are "natural boundaries" for the function. It is not possible to analytically continue the function—to even define it—beyond these walls of singularity.

Later research has shown that this function $\Xi(x)$ is a fractal; as the length of the curve is measured with smaller and smaller line segments, the arclength grows faster than the reciprocal of the length of the segments so that the curve has a fractional dimension. Nevertheless, the Fourier series of $\Xi(x)$ converges geometrically as shown explicitly in (7.17).

On the real axis, $\omega(x;p)$ is a rather bland and harmless-looking function: its fractal jaggedness is evident only along the lines $\operatorname{Im}(x) = \pm a$. Clearly, exponential convergence and a smooth graph on the real axis do not preclude hidden depths of fractal complexity.

The fact that $\omega(x;p)$ is singular *everywhere* on the lines that bound the strip of Fourier convergence implies that the interpolation error for (7.16) will not exhibit strong peaks, but will instead be distributed more or less uniformly over the whole interval like the function itself as shown in Fig. 11.

The function $\omega(x; p)$ is a contrived example. However, Takaoka (1989) has shown that the solitary wave of a fifth-degree, generalized Korteweg–deVries equation is an analytic function with natural boundaries. The natural boundaries are not straight lines like those of $\mathrm{Im}(x)$, but rather are fractal curves which vaguely resemble the letter "v" in the complex x plane. Because the apex of each "v" is a finite distance from the real axis, spectral series for the soliton and for its spatially periodic analogue ("cnoidal wave") converge exponentially fast (Boyd, 1986).

An even more important example (perhaps!) is hydrodynamic turbulence. Frisch *et al.* (1978) and Frisch and Morf (1981) conjecture that the singularities for turbulent flow may form a fractal set in the complex time plane.

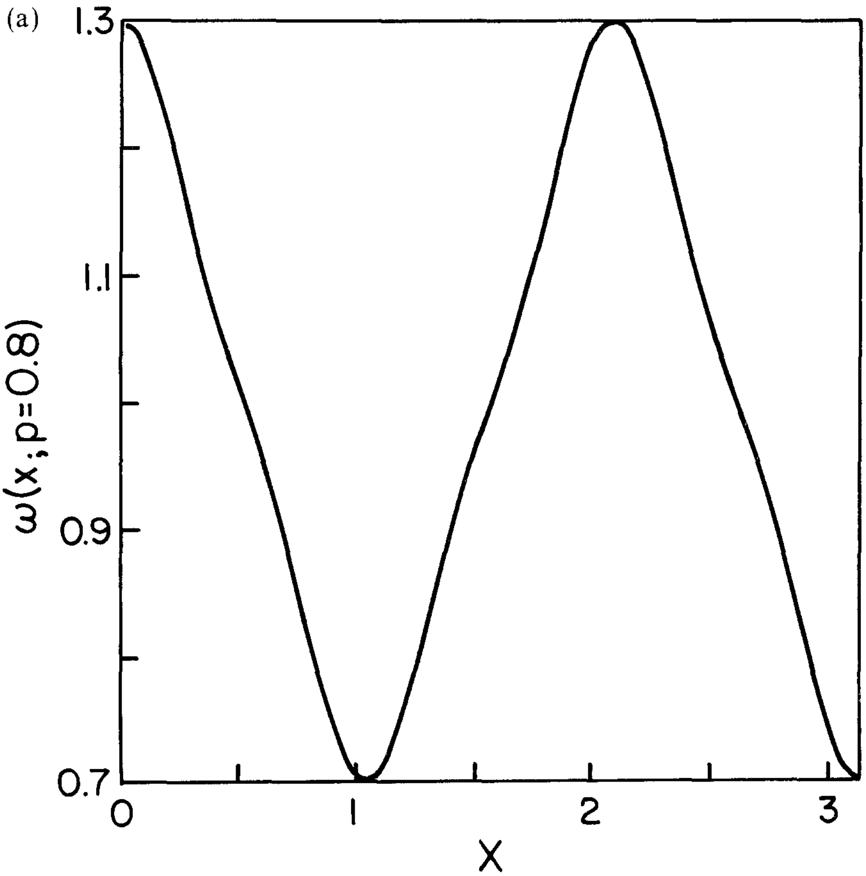

Fig. 11. The Weierstrass-like example, $\omega(x; p)$. (a) $\omega(x; p = 0.8)$.

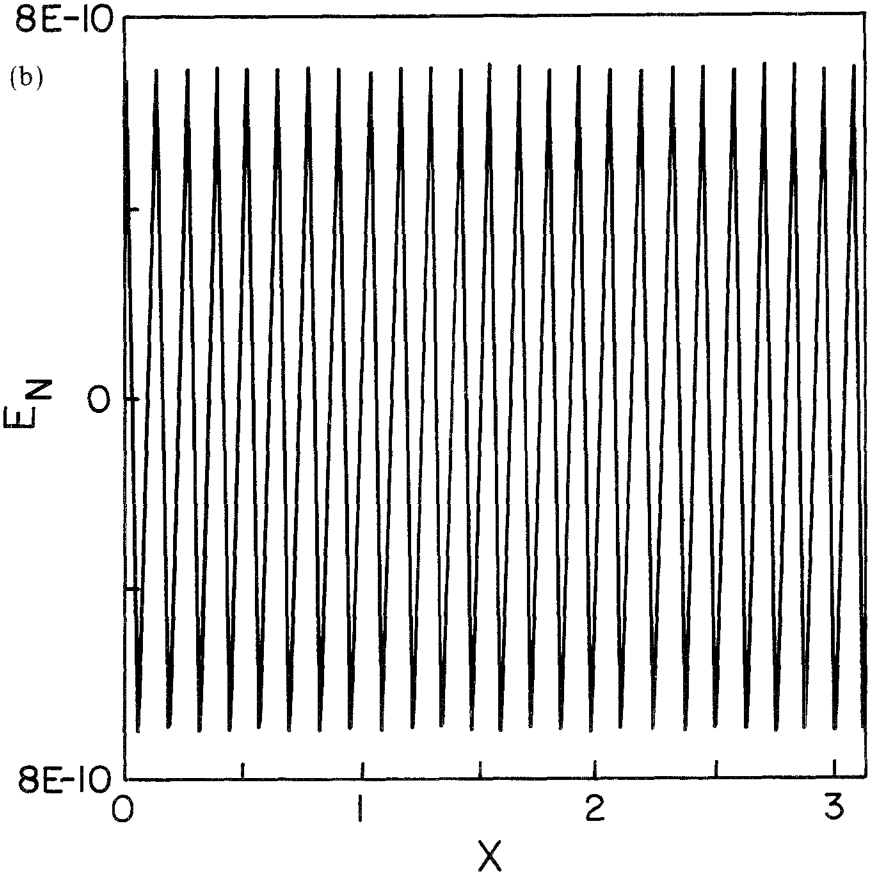

Fig. 11. Continued. (b) interpolation error, $E_I(x;\ N=32)$.

Clearly, functions with natural boundaries are a genuine part of the scientific bestiary. So, too, are functions dominated by a single singularity; even within the narrow realm of solitary waves, the soliton of the ordinary Korteweg–deVries equation has simple second-order poles on the imaginary axis and no other singularities. Its periodic generalization can be approximated by the model function $\lambda_2(x; p)$ described earlier. For nonlinear problems, it is difficult to distinguish between these two extremes—the simplicity of simple poles and the complexity of fractal natural boundaries. The ordinary Korteweg–deVries equation and its fifth degree generalization differ only by the replacement of a third derivative by a fifth derivative.

Nevertheless, our models have at least illustrated the range of possibilities.

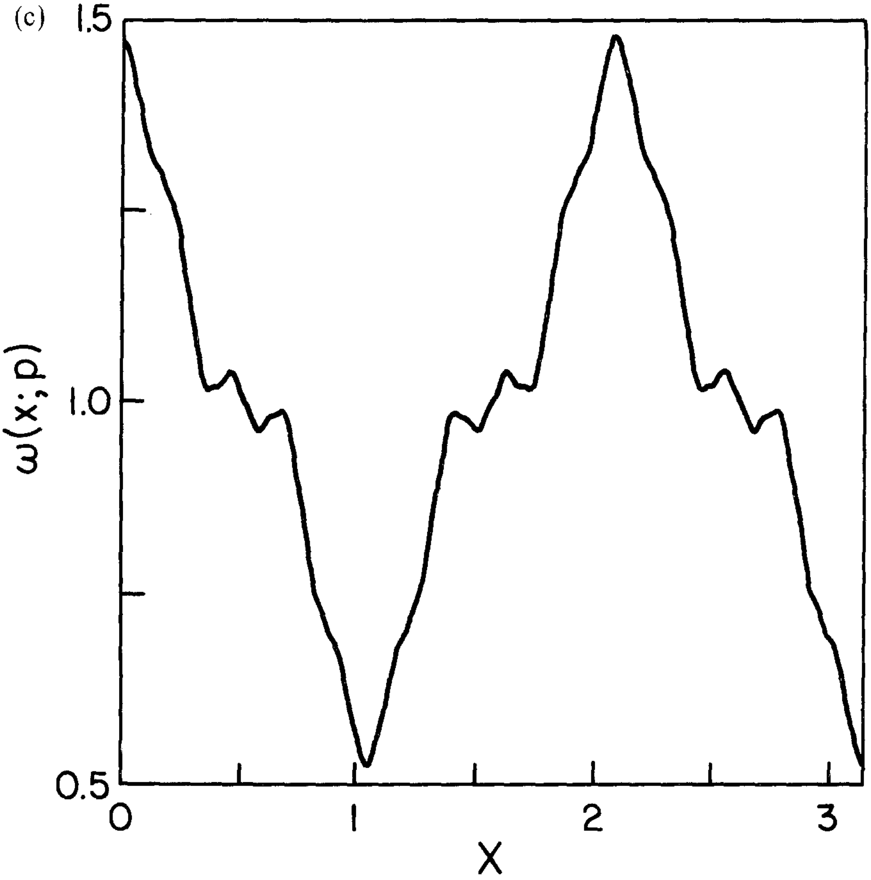

Fig. 11. Continued. (c) $\omega(x; p=0.98)$

8. THE ERROR IN PSEUDOSPECTRAL SOLUTIONS TO DIFFERENTIAL EQUATIONS

To illustrate the key ideas, consider the model

$$u_{xx} + q(x)u = f(x) \tag{8.1}$$

where subscript x denotes differentiation with respect to x and where we assume boundary conditions that $u(x)$ should be periodic with period 2π. For simplicity, we also assume that $q(x)$, $f(x)$, and $u(x)$ are all symmetric about the origin so that the solution is a Fourier cosine series (as opposed to a general Fourier series).

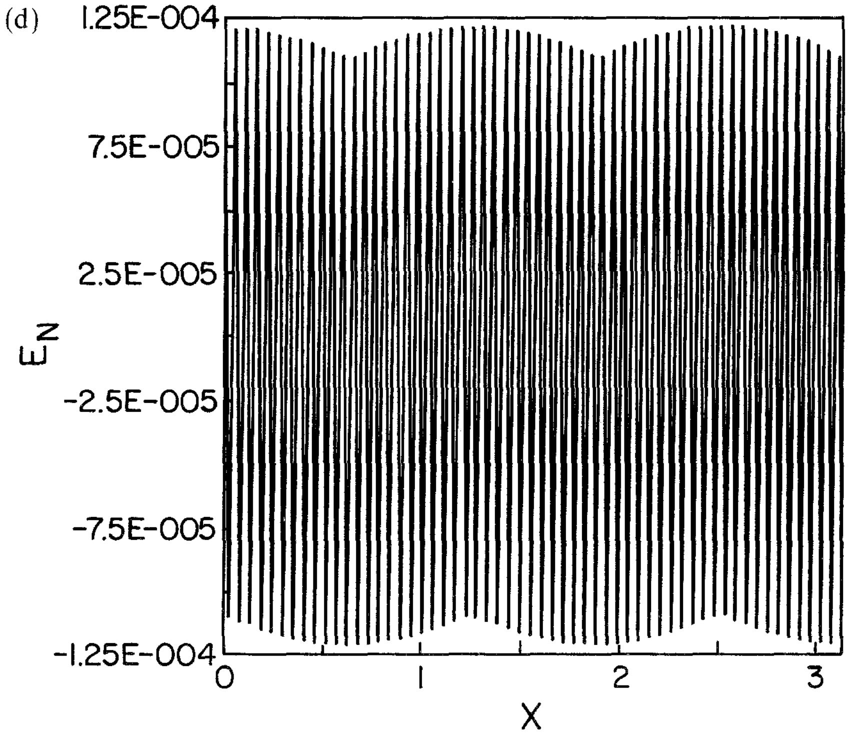

Fig. 11. Continued. (d) $E_I(x;\, N=128)$ when $p=0.98$.

In the pseudospectral method, we write

$$u_N(x) = \sum_{j=0}^{N-1} a_j \cos(jx) \tag{8.2}$$

Substituting this into (8.1) defines the "residual" function

$$R(x;\, N) \equiv f(x) - [u_{N,xx} + q(x)\, u_N(x)] \tag{8.3}$$

The residual function would be identically equal to zero if $u_N(x)$ were the exact solution, but this is too much to hope for except in special cases. Instead, we choose the spectral coefficients $\{a_j\}$ so that the residual function is as small as possible in some sense. In the pseudospectral algorithm, "as small as possible" means that the residual function is forced to be the interpolant of zero. The N unknown spectral coefficients $\{a_j\}$ are then determined by the requirement that $R(x;\, N)$ should vanish at the N interpolation points.

This implies that all the earlier analysis about the envelope of the interpolation error is immediately and directly applicable to the pseudospectral residual function, $R(x; N)$. Since the function being interpolated is $f(x) \equiv 0$, it follows that all of $R(x; N)$ is the interpolation error, $E_I(x; N)$. Thus, the residual has the same qualitative form as all the other examples of $E_I(x; N)$: a slowly varying "envelope" $\rho(x; N)$ multiplying the rapidly oscillating factor $\cos(Nx)$.

Unfortunately, the residual is of only secondary interest because what we really want to know is: What is the error $E_N(x)$ in the pseudospectral approximation to the differential equation? If we define

$$E_N(x) \equiv u(x) - u_N(x) \qquad \text{(pseudospectral error)} \tag{8.4}$$

then subtracting (8.3) from (8.1) shows that

$$E_{N,xx} + q(x) E_N = R(x; N) \tag{8.5}$$

This result is true for any linear differential equation: The pseudospectral error satisfies the differential equation with the substitution $f(x) \to R(x; N)$.

To approximately solve (8.5), define

$$E_N(x) \equiv \sum_{j=0}^{\infty} \varepsilon_j \cos(jx) \tag{8.6}$$

$$R(x; N) \equiv \sum_{j=0}^{\infty} r_j \cos(jx) \tag{8.7}$$

Then for the special case $q(x) \equiv -1$, one finds by matching coefficients of $\cos(jx)$ that

$$\varepsilon_j = -r_j/(1 + j^2) \qquad [q(x) \equiv -1] \tag{8.8}$$

If the forcing function $f(x)$ has an exponentially convergent Fourier series, then the "coefficient-pairing" approximation will apply to the residual $R(x; N)$. This implies that r_N is the largest of the residual coefficients; the magnitude of the coefficients will decrease exponentially fast as the degree increases towards infinity or decreases toward zero.

The high-degree coefficients of the pseudospectral error will behave similarly to those of the residual since the factor of $1/(1 + j^2)$ will simply make the Fourier terms in $E_N(x)$ decrease faster as $j \to \infty$. However, for the low-degree coefficients, there will be competition: The r_j will increase with j until the maximum at $j = N$, whereas the factor of $1/(1 + j^2)$ will decrease with j. In other words, the error coefficients for $j < N$ will be the product of a factor that increases with j (the residual coefficient r_j) and a factor that decreases with j $[1/(1+j^2)]$. Thus, ε_0 will be larger than ε_1 unless r_1 is more than double the magnitude of r_0.

When the magnitudes of the residual coefficients are graphed on a logarithmic scale, the graph will resemble an upside-down "v" with the apex of the "v" at $j=N$. The graph for the error coefficients will be similar except for two differences. First, the apex of the "v" will be lower by $O(N^2)$. Second, there will (sometimes) be a little upward curl for small degree j.

Figure 12a confirms these expectations. The lower panel, Fig. 12b, illustrates how the error coefficients vary with N: the little curl on the left is smallest when N is large, but becomes more pronounced as the number of interpolation points decreases.

To show that these qualitative conclusions are not sensitive to the

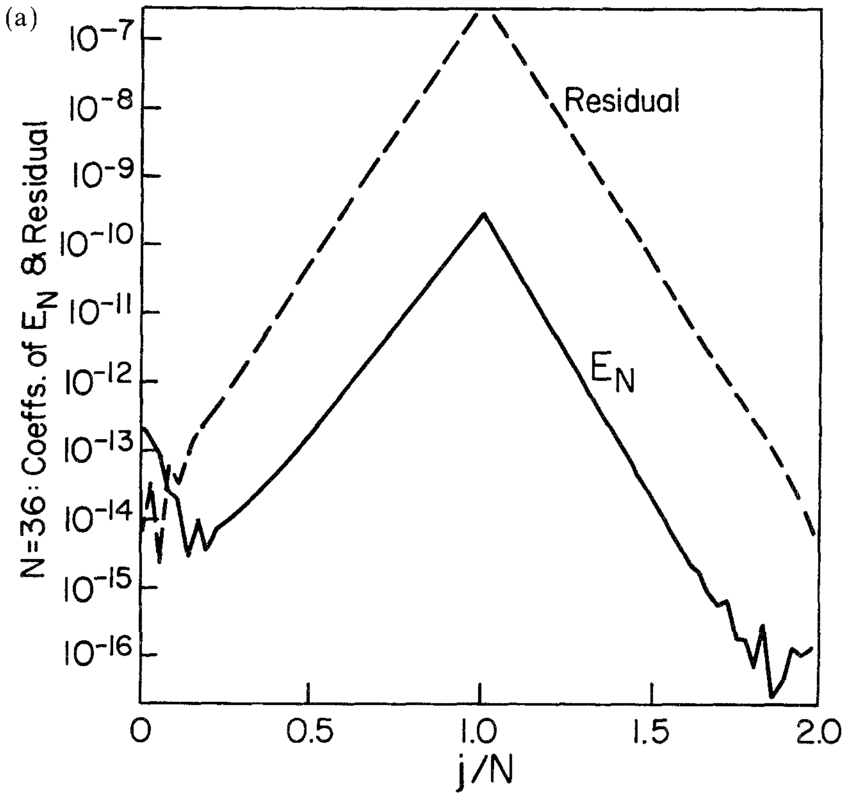

Fig. 12. (a) A comparison of the absolute value of the coefficients of the differential equation residual, $R(x;N)$, with those of $E_N(x;N)$ for $N=36$ and the differential equation (8.9) with $p=0.6$. (b) The absolute values of the spectral coefficients of the error, $E_N(x)$, for three different values of N: top ($N=12$, middle ($N=24$), bottom ($N=36$).

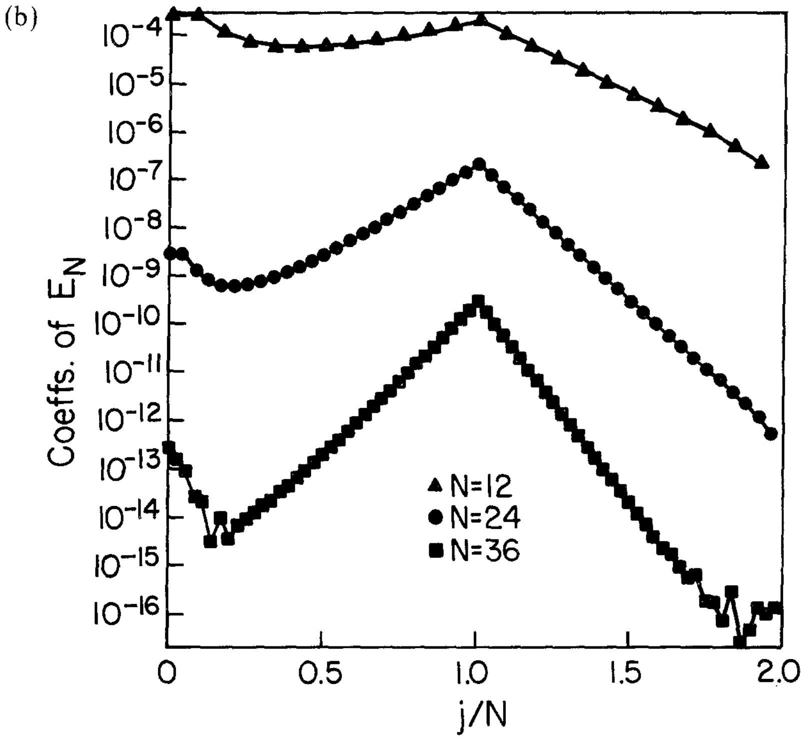

Figure 12. Continued

precise form of the differential equation, Fig. 12 illustrates the pseudospectral solution to the variable coefficient equation,

$$u_{xx}+[-1+\lambda(x;\,p)^2/10]u=\lambda(x;\,p) \tag{8.9}$$

The high-degree coefficients of the error will be dominated by the second derivative, which is $O(j^2)$, as long as $q(x)$ is $O(1)$ even when this function varies with x. The reason is that

$$\begin{aligned}\left[\frac{d^2}{dx^2}+q(x)\right]\cos(jx)&=[-j^2+q(x)]\cos(jx)\\&\approx -j^2\cos(jx)+O(1)\qquad\text{if } j\gg\max|q|\end{aligned} \tag{8.10}$$

Thus, we can ignore variable $q(x)$ in the same way that we approximated $1/(j^2+1)$ by $1/j^2$ in (8.8). It is *generic* that the error coefficients e_j will be roughly proportional to r_j/j^2 except for the low-degree coefficients $[j \sim O(1)]$.

This close relationship between the coefficients of the error $E_N(x)$ and the residual function $R(x; N)$ has several consequences. First, the goal of all reasonable methods for solving differential equations is to minimize the residual. It follows that one of the best strategies for *checking* a calculation is to evaluate the residual and verify that it is small. The analysis of this section shows, however, that the magnitude of the residual is a very *pessimistic* strategy for estimating the error in $u_N(x)$. The reason is expressed by the following theorem.

Theorem 11: *Relative Magnitude of the Residual and Error.* In the limit $N \to \infty$ for fixed $q(x)$ and $f(x)$, the error in solving $u_{xx} + q(x)u = f(x)$ via the pseudospectral method with N interpolation points and a Fourier cosine basis is related to the residual function via

$$\max_{\text{all } x} |E_N(x)| \sim (1/N^2) \max_{\text{all } x} |R(x; N)| \tag{8.11}$$

In this same limit, the largest coefficients in both the error and residual are those of degree N. These are individually proportional with a proportionality constant of N^2:

$$\varepsilon_j \sim -r_j/N^2, \qquad j \approx N \tag{8.12}$$

Proof. We have already explained why the dominant coefficients in both the error and residual are those with $j \approx N$ as illustrated in Fig. 12. Equation (8.8) then immediately implies (8.12). The approximation $j^2 \approx N^2$, which is implicit in (8.12), is obviously restricted to the neighborhood of $j = N$. As shown in Fig. 12, however, both r_j and e_j decrease exponentially fast as $|j - N|$ increases so (8.12) fails only for coefficients that are too small to contribute sufficiently to the maximum of the residual or error. Therefore, the residual as a whole is (approximately) a factor of N^2 larger than the error, justfying the first half of the theorem.

Theorem 11 implies that if $\max[R(x; N)] \sim O(1)$ when $N = 30$, for example, $\max[E_N(x; N)]$ will be only $O(1/900)$. Thus, it is necessary to solve the differential equation to very high accuracy in order to make the residual small. A solution may have less than 1% error and yet still have a large, $O(1)$ residual.

Notwithstanding this caveat, evaluating the residual—perhaps by finite differences with a tiny grid spacing to bypass programming errors in the spectral computation of derivatives—is still highly recommended as a check. The point is simply that this is a very conservative check; a large residual does not necessarily imply an inaccurate solution.

The second consequence of the phenomenology of the error is the perturbative, error correction scheme, which is the theme of the next section.

9. THE METHOD OF MULTIPLE SCALES

The rapid spatial variations of the residual function, which is proportional to $\cos(Nx)$, suggests that we can approximately solve the differential equation satisfied by the error $E_N(x)$, (8.5), by applying the method of multiple scales. The "fast" variable is

$$X \equiv Nx \tag{9.1}$$

while the "slow" variable is x. Rewriting (8.1) and the residual $R(x)$ in terms of the "fast" and "slow" variables gives

$$N^2 E_{N,XX} + q(x) E_N = \cos(X)\, \rho(x; N) \tag{9.2}$$

Neglecting all but the lowest-order terms in N^2 gives the approximate solution

$$E^{(0)} = -R(x)/N^2 \tag{9.3}$$

This is a restatement of Theorem 11: the error is approximately equal to the residual divided by N^2, where N is the number of interpolation points used to generate the pseudospectral solution.

We can iterate by substituting the kth-order approximation into (9.2) and evaluating the residual:

$$R^{(k)}(x; N) \equiv R(x) - E^{(k)}_{xx} - q(x) E^{(k)} \tag{9.4}$$

The refined approximation is

$$E^{(k+1)}(x; N) \equiv -R^{(k)}(x; N)/N^2 + E^{(k)}(x; N) \tag{9.5}$$

The pseudospectral implementation of (9.4) is to evaluate $E^{(k)}$ and $E^{(k)}_{xx}$ on a grid with more than N points—we use $2N$ grid points in Fig. 13—take the Fourier transform to compute the corresponding spectral coefficients, and then differentiate the Fourier series to evaluate the derivatives in (9.4). Another fast cosine transform gives the Fourier coefficients of $R^{(k)}$, and then (9.5) trivially gives the corresponding coefficients of $E^{(k+1)}$.

Figure 13 shows the success—and failure—of this multiple scales iteration for a representative case. The low-wave-number components of the residual do not vary on the "fast" scale, so it is not surprising that the

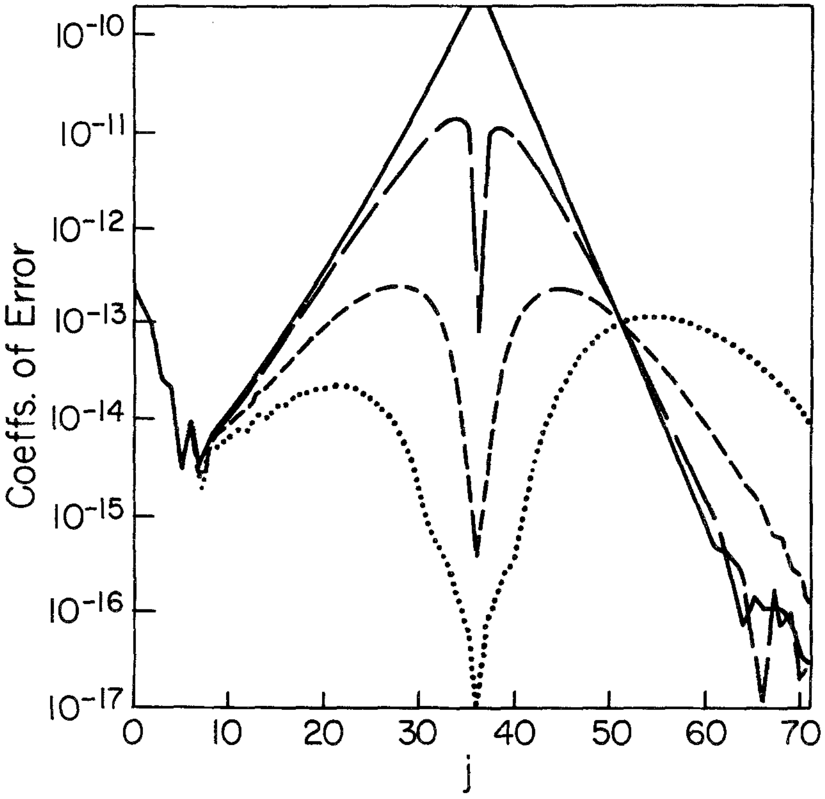

Fig. 13. Errors in the Fourier cosine coefficients for the solution of the differential equation $u_{xx} + [-1 + \lambda(x;\ p = 0.6)^2/10]u = \lambda(x;\ p = 0.6)$. The abscissa is the degree j of the Fourier coefficient. Solid curve: errors in pseudospectral solution with 36 interpolation points (i.e., $N = 36$). Long dashes: errors is zeroth-order multiple scales solution. Short dashes: errors in $u_N(x) + E^{(3)}$. Dotted curve: errors in seventh-order iteration.

iteration fails for this part of the spectrum. The constant in $E_N(x; N)$ should be the same order of magnitude as the constant in $R(x; N)$, but instead (9.4) divides all wave numbers by the huge factor N^2. Consequently, the low-wave-number components of the N-point pseudospectral solution are almost unaffected by the tiny iterative corrections to them.

The middle part of the spectrum, i.e., those components in $E_N(x)$ that are proportional to $\cos(jx)$, where $j \approx N$, are rapidly reduced by the iteration. The only approximation in processing the $j = N$ component is the neglect of $q(x)e_N \cos(Nx)$ relative to the second derivative term, $-N^2 e_N \cos(Nx)$. As a result, the errors in the Fourier components of $E^{(k)}$ show a cusp with the minimum at $j = N$.

For the neighboring components, there is an additional implicit approximation

$$[\cos(jx)]_{xx} \approx -N^2 \cos(jx), \qquad j \approx N \tag{9.6}$$

instead of $-j^2 \cos(jx)$, which is the correct value. Even so, the error in the middle part of the spectrum is rapidly reduced. In consequence, when $E^{(k)}$ is added to $u_N(x)$, the maximum *pointwise* ("L_∞") error decreases rapidly as k increases to moderate values.

Unfortunately, the error grows for the high-wave-number part of the spectrum because (9.6) is a terrible (and unstable) approximation for large j. The result is that the error in the iteration behaves like that for an asymptotic series: It first decreases with k, levels off, and then increases as $k \rightarrow \infty$ as shown in Table IV.

This high-wave-number divergence can be fixed by modifying (9.5) so that the jth wave number is divided by j^2 instead of N^2. We have not used this fix because (9.5) is conceptually simpler and the fix does not solve a more fundamental problem: The low-wave-number components of the error and residual violate the fundamental assumption of varying on the "fast" scale. There is no way to compute the low-wave-number components of the error from those of the residual without performing additional matrix solves. Such convergent-but-matrix-solving iterations are the theme of the companion paper, Boyd (1991), so we shall not discuss them here.

Table IV. Error Reduction in the Multiple Scales Series[a]

Multiple scales order	L_∞ Error $\times 10^9$)
Pseudospectral	1.29
0	0.0936
1	0.0344
2	0.00852
3	0.00554
4	0.00228
5	0.00240
6	0.00176
7	0.00234

[a] The first entry is the error in the 36-point pseudospectral solution; the second row (zeroth order) is the result of correcting the pseudospectral solution by subtracting $E^{(0)}(x)$, and similarly for higher orders.

It is nevertheless remarkable that at the cost of only a few Fourier transforms, we can remove the spectral peak of the error in the N-point pseudospectral solution and lower it to the much smaller magnitude of the error in the lowest few wave numbers. The success of the iteration, even the limited success of an asymptotic rather than a convergent approximation, confirms the validity of the multiple scales concept. To a good approximation when N is sufficiently large, the error in the solution to a differential equation has the envelope-times-$\cos(Nx)$ structure of interpolation errors.

10. CHEBYSHEV POLYNOMIALS

To apply the concepts discussed above for *interpolation* to *Chebyshev polynomials*, the only changes are that there are no changes. The reason is that a Chebyshev series is merely a cosine series in disguise as expressed by the identity

$$T_n(\cos[t]) \equiv \cos(nt) \tag{10.1}$$

for all n.

Under this change of variable

$$x = \cos(t) \tag{10.2}$$

the singularity type is unaltered except at the end points, $x = \pm 1$. Thus, the model function $\lambda(t;\, p)$, which is a model of a function with simple poles in the complex plane, is transformed by the mapping into the function $\lambda(\arccos(x);\, p)$, which still has a complex conjugate pair of simple poles on the imaginary axis.

The error for Chebyshev interpolation is a slowly varying factor $\rho(x;\, N)$ muliplied by a rapidly oscillating factor; the only modification is that the "fast" factor is $T_N(x)$. This oscillates between 1 and -1 just like $\cos(Nx)$, but $T_N(x)$ varies more rapidly near the end points $x = \pm 1$ than near the middle, whereas its Fourier counterpart, $\cos(Nx)$, oscillates with uniform frequency for all x.

The change of variable (10.2) maps lines parallel to the real t axis into ellipses with foci at ± 1 in the complex x plane. We noted earlier that in the Fourier case, shifting a singularity along a line parallel to the real axis has no significant efect on the rate of Fourier convergence; such shifts induce oscillations in the Fourier coefficients, but not in the rate at which the coefficients a_n decrease as $n \to \infty$. For Chebyshev series, the equivalent statement is that the rate at which coefficients decrease is independent of location along a particular ellipse, but does depend on the size of the ellipse.

Thus, for interpolation, the differences between the Fourier and Chebyshev cases are mostly matters of notation and interpretation. The concept of the envelope of the interpolation error is unchanged.

The Fourier versus Chebyshev differences are considerably greater for *differential equations*, i.e., the theme of the preceding section. The reason is boundary conditions. When the cosines are the basis, the boundary condition is periodicity, which is automatically satisfied by each basis function. When Chebyshev polynomials are used, the usual boundary conditions are Dirichlet conditions such as

$$u(-1) = u(1) = 0 \tag{10.3}$$

We must either add rows to the pseudospectral matrix to explicitly impose these conditions, or modify the basis set.

Nevertheless, much of the analysis of the Fourier pseudospectral method can be extended to the Chebyshev case if we adopt basis modification by writing

$$u_N(x) \equiv (1 - x^2) \sum_{j=0}^{\infty} a_j T_j(x) \tag{10.4}$$

so that the approximate solution exactly satisfies the homogeneous boundary conditions (10.3). [We can generalize to inhomogeneous boundary conditions by writing $u(x) = v(x) + B(x)$, where $B(x)$ satisfies the boundary conditions, and then using (10.4) to compute an approximation to $v(x)$.] The advantage of choosing this form is that most of the Fourier multiple scales analysis carries over to the Chebyshev pseudospectral method, too.

First, note that the pseudospectral method chooses the approximate spectral coefficients so that the residual of the differential equation is the interpolant of zero. This implies that the Chebyshev residual must have the form

$$R(x; N) = \rho(x; N)\, T_N(x) \tag{10.5}$$

where $\rho(x; N)$ is the envelope. The "coefficient-pairing" approximation of Sec. 4 applies without modification; thus the residual can be written

$$R(x; N) \approx r_N T_N(x) + \sum_{k=1}^{N} r_{N+k} [T_{N+k}(x) + T_{N-k}(x)] \tag{10.6}$$

This implies that the Chebyshev coefficients of $R(x; N)$ will have the upside down "v" shape of the Fourier coefficients of the residual: r_j will be strongly peaked at $j = N$.

The complication is that whereas the second derivative of $\cos(jx)$ is proportional to $\cos(jx)$, the second derivative of a Chebyshev polynomial is messy. However, by using (10.1) and (10.2), it is easy to show that

$$[(1-x^2)T_j]_{xx} = -j^2\cos(jt) - 3j\cos(t)\sin(jt)/\sin(t) - 2\cos(jt) \qquad (10.7)$$

$$\sim -j^2\cos(jt)[1+O(1/j)] \qquad (10.8)$$

which, using $\sin^2(t) = (1-x^2)$, is equivalent to

$$[(1-x^2)T_j]_{xx} \sim -j^2 T_j(x)[1+O(1/j)] \qquad (10.9)$$

It follows from (10.9) that the rest of the multiple scales analysis applies just as for Fourier series. If the differential equation is

$$u_{xx} + q(x)u = f(x) \qquad (10.10)$$

then the error $R_N(x)$ satisfies precisely the same equation as its Fourier counterpart:

$$E_{N,xx} + q(x)E_N = R(x; N) \qquad (8.5')$$

Because the pseudospectral solution satisfies the boundary conditions exactly, the error at the boundaries is zero for all N. It follows that we can write

$$E_N(x) \equiv (1-x^2)\sum_{j=0}^{\infty} \varepsilon_j T_j(x) \qquad (10.11)$$

Then one finds that the approximation (8.12) also holds:

$$\varepsilon_j \sim -r_j/N^2, \qquad j \approx N \qquad (10.12)$$

which in turn implies that

$$E_N(x) \sim -R(x)/N^2 \qquad \text{as} \quad N \to \infty \qquad (10.13)$$

which is a restatement of (9.3).

We omit a detailed (and boring) repetition of the rest of Secs. 8 and 9 because it should be clear that the Chebyshev and Fourier analysis is very similar. In Boyd (1991), we extend the present analysis by deriving a convergent iteration based on the Chebyshev multiple scales analysis.

11. SUMMARY

In this work, we have derived explicit expressions for the interpolation error and also the error in pseudospectral solutions to differential

Table V. A Summary of Results

(a) Basic definitions

$$f(x) \equiv \text{function being interpolated} \equiv \alpha_0/2 + \sum_{j=1}^{\infty} \alpha_j \cos(jx)$$

$$\rho(x;N) \equiv \text{envelope of interpolation error} \equiv \frac{f(x)}{\cos(Nx)} = \rho_0/2 + \sum_{j=1}^{\infty} \rho_j \cos(jx)$$

$$E_T(x;N) \equiv \text{truncation error} \equiv f(x) - P_N f \equiv \sum_{j=N}^{\infty} e_j^{(T)} \cos(jx) \equiv \sum_{j=N}^{\infty} \alpha_j \cos(jx)$$

$$E_I(x;N) \equiv \text{interpolation error} \equiv f(x) - I_N f = \sum_{j=1}^{\infty} e_j^{(I)} \cos(jx) + e_0^{(I)}/2$$

(b) Theorems

1. $e_j^{(T)} = e_j^{(I)} = \alpha_j, \quad j \geqslant N$ [EXACT]

2. $\rho_j = \sum_{m=0}^{\infty} (-1)^m \alpha_{j+(2m+1)N}, \quad j = 0, 1, \ldots$ [EXACT]

3. If $f(x) = \lambda(x;p) = 1 + 2\sum_{j=1}^{\infty} p^j \cos(jx)$, then the errors are [EXACT]:

$$E_I = \left(\frac{2p^N}{1+p^{2N}}\right)\cos(Nx)\,\lambda(x;p) \quad \text{and} \quad E_T = \left(\frac{2p^N}{1-p^2}\right)\{\cos(Nx) - p\cos([N-1]x)\}\,\lambda(x;p)$$

If $f(x) = \mu(x;p) \equiv 2\sum_{j=1}^{\infty} p^j \sin(jx)$ then the errors are [EXACT]

$$E_I = \left(\frac{2p^N}{1+p^{2N}}\right)\cos(Nx)\,\mu(x;p) \quad \text{and} \quad E_T = -2p^{N+1}\frac{\{p\sin(Nx) - \sin([N+1]x)\}}{(1+p^2) - 2p\cos(x)}$$

4. NTA ("neglect-of-triple-aliasing") approximation:

$$\rho_j = 2\alpha_{j+N} + O(\alpha_{j+3N}), \qquad j = 0, 1, \ldots$$

5. Aliasing relations. Let the interpolation grid be

$$x_i \equiv \pi(2i-1)/(2N), \qquad i = 1, 2, K, N$$

Then, letting the symbol $\overset{G}{=}$ denote an equality that holds only at the points of the interpolation grid (and not for intermediate x), we have

$$\cos(kx + 2Nmx) \overset{G}{=} (-1)^m \cos(kx) \qquad [k, m \text{ integers}]\ \sin(kx + 2Nmx) \overset{G}{=} (-1)^m \sin(kx)$$

6. Low degree coefficients for E_I

$$e_k^{(I)} = \sum_{m=1}^{\infty} (-1)^{m+1}\{\alpha_{2Nm-k} + \alpha_{2Nm+k}\}, \qquad k = 0, 1, \ldots, N-1 \quad [\text{EXACT}]$$

Table continued

Table V (*Continued*)

7. Coefficient pairing (neglect-of-double-aliasing) approximation:

$$E_I(x; N) = \alpha_N \cos(Nx) + \sum_{j=1}^{N} \alpha_{N+j}\{\cos[(N+j)x] + \cos[(N-j)x]\} + O(\alpha_{2N+1})$$

8. "Effective Lorentzian" approximation:

$$\rho(x; N) \approx \alpha_N \, \lambda(x; p_{\mathrm{eff}}), \qquad \text{where} \quad p_{\mathrm{eff}} \equiv e^{\alpha'(N)/\alpha(N)}$$

where the prime denotes the derivative of $a(j)$ with respect to degree. The error is $O(e^{j^2[\alpha'(N)/\alpha(N)]})$.

9. Interpolation error for phase-shifted lorentzians. Let ρ_λ and ρ_μ denote the envelope of the N-point interpolation error to the Lorentzian functions $\lambda(x+\phi; p)$ and $\mu(x+\phi; p)$ (defined in Theorem III]. In the NTA approximation, these are given with a relative error of $O(p^{2N})$ by

$$\left|\begin{matrix} \rho_\lambda(x; p, N) \\ \rho_\mu(x; p, N) \end{matrix}\right| = 2p^N \left|\begin{matrix} \cos(N\phi) & -\sin(N\phi) \\ \sin(N\phi) & \cos(N\phi) \end{matrix}\right| \left|\begin{matrix} \lambda(x+\phi; p) \\ \mu(x+\phi; p) \end{matrix}\right|$$

10. Interpolation error for the models $\lambda_k(x; p)$ in the NTA approximation. Let the model functions be defined by

$$\lambda_k(x; p) \equiv 2 \sum_{j=1}^{\infty} j^k p^j \cos(jx)$$

Then we have

$$E_I(x; N) \approx 2N^k p^N \cos(Nx) \left[\lambda(x; p) + \sum_{m=1}^{\infty} \frac{\Gamma(m+1)}{\Gamma(m+1)\,\Gamma(k-m+1) N^m} \lambda_m(x; p) \right]$$

where the series terminates at $m = k$ if k is a positive integer.

11. Relative magnitude of the differential equation residual and error. The error E_N in the N-point collocation solution to a second-order differential equation is related to the residual function R as

$$\max|E_N| \sim \frac{1}{N^2} \max|R(x; N)| \qquad \text{as} \quad N \to \infty$$

equations. For simplicity, most theorems are given in terms of Fourier series, but as explained in Sec. 10, almost all results apply with only trivial modifications to Chebyshev polynomials. In both cases, the interpolation error takes the form of a rapidly varying factor, $\cos(Nx)$ or $T_N(x)$, multiplying a slowly varying factor or "envelope."

Table V is a summary of theorems. Both exact and approximate expressions for the error are given because the approximate formulas are simpler and easier to interpret (at the expense of accuracy).

One striking conclusion is that for simple functions, i.e., those with a single dominant pair of complex conjugate singularities or a single peak, the Fourier/Chebyshev error is anything but uniform. Instead, the error envelope is sharply peaked about the point where $f(x)$ is peaked or which is nearest the poles or branch points.

Typically, the envelope can be approximated by a linear combination of the "Lorentzian" function $\lambda(x+\phi; p)$ and the "serpentine" $\mu(x+\phi; p)$ defined in Theorem 3 and illustrated in Figs. 2 and 3. The phase shift constant ϕ shifts the peaks of the Lorentzian and serpentine so that $\lambda(x+\phi;p)$ and $\mu(x+\phi; p)$ are centered on the singularities of $f(x)$ [or the crests and troughs of $f(x)$ if it is an entire function]. The parameter p, which measures how sharply the error is peaked, depends on the type and strength of the singularities of $f(x)$ and the distance of these poles or branch points from the expansion interval.

Another striking conclusion is that the error in solving a second-order differential equation using the pseudospectral method with N collocation points is typically smaller than the residual $R(x; N)$ by a factor of $O(N^2)$. [$O(N^4)$ for a fourth-order equation.] [We have to hide behind the weasel word "typically" because if one of the low-degree basis functions is close to being a zero-eigenvalue eigenmode of the differential equation, then the error could be dominated by that single component, invalidating (10.13).] Excluding this rare exception, checking a numerical solution by evaluating the residual is a very conservative check: $R(x)$ may be $O(1)$ even when the error $E_N \sim O(1/N^2) \ll 1$.

Nevertheless, checking via $R(x)$ is still a very valuable verification tool. Section 9 shows merely that this tool must be used carefully.

ACKNOWLEDGMENTS

This work was suported by the NSF through grants Nos. DMS8716766 and OCE8812300. I thank Springer-Verlag for permission to reprint some figures from the author's monograph, Boyd (1989a).

REFERENCES

Boyd, J. P. (1986). Solitons from sine waves: Analytical and numerical methods for non-integrable solitary and cnoidal waves, *Physica* **21D**, 227–246.

Boyd, J. P. (1989a). *Chebyshev and Fourier Spectral Methods*, Springer-Verlag, New York.

Boyd, J. P. (1989b). New directions in solitons and nonlinear periodic waves: Polycnoidal waves, imbricated solitons, weakly nonlocal solitary waves, and numerical boundary value algorithms, in *Advances in Applied Mechanics*, Vol. 27, Wu, T.-Y., and Hutchinson J. W. (eds.), Academic Press, New York, pp. 1–82.

Boyd, J. P. (1991). The Delves–Freeman–Lanczos iteration. To be published.

Elliott, D. (1965). Truncation errors in two Chebyshev series approximations, *Math. Comput.* **19**, 234–248.

Frisch, U., Sulem, P.-L., and Nelkin, M. (1978). *J. Fluid Mech.* **87**, 719.

Frisch, U., and Morf, R. (1981). Intermittency in nonlinear dynamics and singularities at complex time, *Phys. Rev. A* **23**, 2673–2705.

Gradshteyn, I. S., and Ryzhik, I. M. (1965). *Tables of Integrals, Series, and Products*, Academic Press, New York.

Rucker, R. (1987). *Mind Tools*. Houghton-Mifflin, New York, p. 135.

Takaoka, M. (1989). Pole distribution and steady pulse solution of the fifth order Korteweg–deVries equation. Preprint, Dept. of Physics, Kyoto University, Kyoto 606.

Voss, R. F. (1988). In, *The Science of Fractal Images*, Peitgen, H.-O., and Saupe, D. (eds.), Springer-Verlag, New York, pp. 21–70.

Printed in Belgium