The Carathéodory-Fejér Method for Recursive Digital Filter Design

MARTIN H. GUTKNECHT, JULIUS O. SMITH, MEMBER, IEEE, AND LLOYD N. TREFETHEN

Abstract—A new technique for rational digital filter design is presented which is based on results in complex function theory due to Takagi, Krein, and others. Starting from a truncated or windowed impulse response, the method computes the unique optimum rational Chebyshev approximation with a prescribed number of stable poles. Both phase and magnitude are matched. Deleting the noncausal (unstable) part of the Chebyshev approximation yields a stable approximation of specified order (M, N) which is close to optimal in the Chebyshev sense. No iteration is involved except in the determination of an eigenvalue and eigenvector of the Hankel matrix of impulse response coefficients. In this paper the algorithm is specified and practical examples are discussed.

I. Introduction

No FAST and reliable algorithm exists for the optimal Chebyshev approximation of an arbitrary magnitude characteristic $|H(e^{j\omega})|$ or frequency response $H(e^{j\omega})$ by a stable infinite-impulse-response (IIR) filter. In principle, one must solve a real nonlinear approximation problem with respect to a weighted Chebyshev norm (when approximating $|H(e^{j\omega})|$) or a complex one [when approximating $H(e^{j\omega})$], but these tasks are very difficult. Although versions of the Remez algorithm [13], nonlinear programming techniques [4], [23], and the differential correction algorithm [5], [15] have been used in the real case, and a descent algorithm [19] and the Lawson algorithm [3], [16] have been tried in the complex case, none of these methods can claim to be fast. In the complex case, no method is known to be globally convergent to an optimum solution from an arbitrary starting point [20], and optimal solutions are in general not unique [47], [48].

Although there exist numerous techniques for near-best approximation [6], [7], [10], [12], [31], [41], resulting designs are not always stable; hence it is often necessary to find and modify the unstable poles of the filter, which is typically much more work than the filter design algorithm itself [12], [13], [31], [41].

We present a new filter design method that allows one to compute a *stable*, near-optimal, uniformly weighted approximation to the complex function $H(e^{j\omega})$. The method is based on an extension due primarily to Takagi [21], [40], of a classical theorem in complex analysis of Carathéodory and Fejér [9]. Consequently, we refer to this technique as the *Carathéodory-Fejér* (CF) algorithm. The CF method is most powerful for approximating smooth functions, but it is also competitive for any filter design problem in which minimization of some norm of the complex frequency-response error is desired.

The CF method requires that the Fourier coefficients of $H(e^{j\omega})$, i.e., its impulse response, be given. We assume the given impulse response is causal and finite. Although the CF method works in the time domain, it delivers an approximation of the frequency response $H(e^{j\omega})$ which is close to optimal in the Chebyshev sense.

An important case in practice is the situation in which only the magnitude of the frequency response $|H(e^{j\omega})|$ is prescribed. For problems of this type we normally generate a complex spectrum having the same magnitude and minimum phase, using homomorphic signal processing techniques [14], [29]. Alternatively, we may compute a high-order finite-impulse-response (FIR) filter to use as a basis of the IIR filter design [30], [33]; and in this case, recursive filters with approximately linear phase may be designed.

Techniques related to the CF method are currently attracting attention in the field of linear systems theory, particularly regarding the *model order reduction* problem. While the CF method is nearly optimal in the Chebyshev norm, it is in fact optimal in the so-called *Hankel norm* when $M \geqslant N - 1$, where M is the number of zeros and N is the number of poles. (The Hankel norm is defined as the spectral norm of the Hankel matrix of the impulse response error.) Presentations of some of the work on model order reduction in the Hankel norm may be found in [17], [18], [26], [27], [39]. However, there seems to be no previous work applying Hankel norm minimization to the filter design problem.

In the model order reduction problem, the starting point is a rational digital filter which is to be approximated by a lower order rational filter. This problem may be adapted to digital filter design by taking an FIR filter as the starting point. However, the CF method is not equivalent to an existing model order reduction method applied to FIR filters. First, we eliminate the usual restriction $M = N - 1$. Second, the above

Manuscript received May 14, 1980; revised June 14, 1982 and May 1, 1983. This work was supported by the Swiss NSF Research Fellowship under Grant 82.723.0.79, the Hertz Foundation, an NSF Graduate Fellowship, and the Office of Naval Research under Contract N00014-75-C-1132.

M. H. Gutknecht was on leave at the Department of Computer Science, Stanford University, Stanford, CA 94305. He is with the Seminar für Angewandte Mathematik, Eidgenössische Technische Hochschule, Zürich, Switzerland.

J. O. Smith was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305. He is now with the Adaptive Systems Department, Systems Control Technology, Palo Alto, CA 94303.

L. N. Trefethen was with the Department of Computer Science, Stanford University, Stanford, CA 94305. He is now with the Courant Institute of Mathematical Sciences, New York University, New York, NY 10012.

0096-3518/83/1200-1417$01.00 © 1983 IEEE

references propose methods which include a partial fraction expansion. Our experience has shown that this can limit the length of the original impulse response which can be used, since this length is the size of the polynomial which must be factored. The CF method circumvents partial fraction expansion by means of an FFT-based spectral factorization technique, applicable whenever the numbers of poles and zeros inside the unit circle are known *a priori*.

II. Theoretical Basis of the CF Method

Assume an ideal causal impulse response $h(n) (n = 0, 1, \cdots)$ is given, corresponding to an ideal transfer function

$$H(z) \triangleq \sum_{n=0}^{\infty} h(n) z^{-n}.$$

If this series converges uniformly on the unit circle, $H(z)$ can be approximated arbitrarily closely by taking a partial sum of sufficiently high order. In practice, it may be preferable to apply a band-limited window [33] rather than truncate, and we denote the possibly modified impulse-response values by $\{h_K(n)\}_0^K$ and the corresponding transfer function by $H_K(z)$

$$H_K(z) \triangleq \sum_{n=0}^{K} h_K(n) z^{-n}. \tag{2.1}$$

We address the problem of approximating $H_K(z)$ on the unit circle $\Gamma \triangleq \{z \in C\colon |z| = 1\}$ by a rational transfer function

$$R_{MN}(z) \triangleq \frac{B(z)}{A(z)} \triangleq \frac{\sum_{k=0}^{M} b_k z^{-k}}{\sum_{k=0}^{N} a_k z^{-k}} \tag{2.2}$$

with all poles inside Γ, and normalized by $a_0 = 1$. We denote by $\mathcal{R}_{MN} (M, N \geqslant 0)$ the set of all such functions.

An optimal (complex) rational approximation to H_K on Γ under the Chebyshev norm is any $R_{MN}^* \in \mathcal{R}_{MN}$ which satisfies

$$\|R_{MN}^* - H_K\|_\infty \leqslant \|R_{MN} - H_K\|_\infty \qquad \text{for all } R_{MN} \in \mathcal{R}_{MN}$$

where $\|f\|_\infty \triangleq \max\{|f(z)|\colon z \in \Gamma\}$. As we have stated, it is in general very difficult to compute such a function R_{MN}^*. It happens, however, that it is easy to determine the best Chebyshev approximation $\tilde{R}_{MN}^*$ out of the larger class $\tilde{\mathcal{R}}_{MN}$ of functions which are of the form

$$\tilde{R}_{MN}(z) \triangleq \frac{\tilde{B}(z)}{A(z)} \triangleq \frac{\sum_{k=-\infty}^{M} b_k z^{-k}}{\sum_{k=0}^{N} a_k z^{-k}} \tag{2.3}$$

(with $a_0 = 1$), where the zeros of $z^N A(z)$ still lie inside Γ and the series in the numerator converges there. The class $\tilde{\mathcal{R}}_{MN}$ may be regarded as an extension of the filters in $\mathcal{R}_{MN}$ to include noncausal impulse response conponents. The CF method consists of computing this extended best approximation $\tilde{R}_{MN}^*$, and truncating it to obtain the CF approximant $R_{MN}^{(\mathrm{CF})} \in \mathcal{R}_{MN}$. One way to perform this truncation is to express $\tilde{R}_{MN}^*$ in the parametric form (2.3) and delete the terms with negative k in the numerator [22]. A better method, which we employ here, is to start with the impulse response (Laurent series) for (2.3). If $M \geqslant N - 1$, then one simply truncates all noncausal terms, and what remains is the impulse response for a function in $\mathcal{R}_{MN}$. For $M < N - 1$, a slight modification of this procedure is necessary, as described in the next section.

It may appear that the filter $R_{MN}^{(\mathrm{CF})}$ obtained by truncating R_{MN}^* as above will in general be far from optimal, but in fact it is exceedingly close to optimal in the Chebyshev sense. One can see this to some extent by considering that if $\tilde{R}_{MN}^*$ fits both the magnitude and phase of the causal transfer function H_K closely in the Chebyshev norm, then $\tilde{R}_{MN}^*$ must itself be approximately causal. In fact, if $\|\tilde{R}_{MN}^* - H_K\|_\infty = \lambda$, then each noncausal term of the impulse response of $\tilde{R}_{MN}^*$ has magnitude at most λ, and the noncausal terms approach zero exponentially in the negative time direction. However, the truncation error is typically a good deal smaller than this. For some estimates on its size, see [42] and [43]. In the case $M \geqslant N - 1$, as we have mentioned, one can show that $R_{MN}^{(\mathrm{CF})}$ in fact approximates H_K optimally in the Hankel norm.

It remains to describe the method for computing $\tilde{R}_{MN}^*$. The answer is given by a theorem developed by Takagi [40], Akhieser [2], Clark [11], and Adamjan *et al.* [1], for which an elementary proof is given in [43]. For a detailed presentation of the Takagi theory, see also [21]. The polynomial case $(N = 0)$ was settled earlier by Carathéodory and Fejér [9].

The theorem makes use of the *singular value decomposition* of the *Hankel matrix* formed from the windowed impulse response $\{h_K(n)\}_{n=0}^K$. The values $h_K(n)$ may be complex. By definition, the Hankel matrix corresponding to an impulse response $\{h_K(n)\}_{n=0}^\infty$ is the infinite matrix having $h_K(i + j)$ at the intersection of the ith row and jth column $(i, j = 0, 1, 2, \cdots)$. To obtain general type (M, N) approximations, we introduce the parameter

$$\nu \triangleq M - N + 1$$

and define the Hankel matrix entry (i, j) as $h_K(i + j + \nu)$,

$$\mathbf{H}_{\nu,K} \triangleq \begin{pmatrix} h_K(\nu) & h_K(\nu+1) & \cdots & h_K(K) \\ h_K(\nu+1) & & \cdot^{\cdot^{\cdot}} & 0 \\ \vdots & h_K(K) & & \vdots \\ h_K(K) & 0 & \cdots & 0 \end{pmatrix} \tag{2.4}$$

where $h_K(k) \triangleq 0$ for $k < 0$.

The singular value decomposition of $\mathbf{H}_{\nu,K}$ may be expressed as

$$\mathbf{H}_{\nu,K} = U \Sigma V^* \tag{2.5}$$

where U, V are unitary matrices, and Σ is a diagonal matrix with nonnegative diagonal elements $\sigma_0, \cdots, \sigma_{K-\nu}$ arranged in order of decreasing magnitude [38]. These elements of Σ are called the *singular values* of $\mathbf{H}_{\nu,K}$. (Note that it is customary to number the singular values from 1 rather than 0. Our choice is made to simplify notation. Also, we refer to σ_n as the nth singular value, although it is the $(n + 1)$st element of the sequence.) The left and right *singular vectors* corresponding to σ_n are the nth columns of U and V, respectively, and we denote them by

$$U_n \triangleq (u_n(0), \cdots, u_n(K-\nu))^T,$$
$$V_n \triangleq (v_n(0), \cdots, v_n(K-\nu))^T.$$

If σ_n is not a simple singular value, then U_n and V_n are not unique, but this does not matter in the theorem below.

When the impulse response is real, $\mathbf{H}_{\nu,K}$ is a real symmetric matrix, and in this case $\sigma_n = |\lambda_n|$, where λ_n is the nth eigenvalue of $\mathbf{H}_{\nu,K}$ by magnitude $(|\lambda_0| \geqslant |\lambda_1| \geqslant \cdots \geqslant |\lambda_{K-\nu}|)$. Moreover, in this case one may assume

$$V_n = U_n \,\mathrm{sign}(\lambda_n). \tag{2.6}$$

Thus, in the case of a real impulse response (i.e., for real symmetric matrices), each singular vector is also an eigenvector and vice versa.

We now quote from [43] the important result on which our method is based (see also [2], [21], [40]).

Theorem: H_K has a unique best Chebyshev approximation $\tilde{R}^*_{MN}$ out of $\tilde{\mathcal{R}}_{MN}$, and the error function $(H_K - \tilde{R}^*_{MN})(z)$ is an all-pass filter having constant modulus on $|z| = 1$. The error modulus is equal to the Nth singular value of the Hankel matrix $\mathbf{H}_{\nu,K}$, i.e.,

$$\|H_K - \tilde{R}^*_{MN}\|_\infty = \sigma_N \tag{2.7}$$

where $\sigma_N \triangleq 0$ for $N > K - \nu$. $\tilde{R}^*_{MN}$ is given by

$$\tilde{R}^*_{MN}(z) = H_K(z) - \sigma_N z^{-\nu} \frac{U_N(z)}{V_N(z^{-1})} \tag{2.8}$$

where $U_N(z)$ and $V_N(z)$ are formed from the Nth singular vectors of $\mathbf{H}_{\nu,K}$ as

$$U_N(z) \triangleq \sum_{n=0}^{K-\nu} u_N(n) z^{-n},$$

$$V_N(z) \triangleq \sum_{n=0}^{K-\nu} v_N(n) z^{-n}.$$

The theorem implies that every stable linear system (of arbitrary order) admits a decomposition into the sum of a noncausal rational filter from the class $\tilde{R}^*_{MN}$ plus an all-pass filter

$$H_K(z) = \tilde{R}^*_{MN}(z) + \sigma_N z^{-\nu} \frac{U_N(z)}{V_N(z^{-1})}.$$

In the proof of the theorem, this equation follows immediately from taking the z-transform of the equation $\mathbf{H}_{\nu,K} V_N = \sigma_N U_N$, which follows from (2.5). What is nontrivial to show, however, is that the number of poles of $\tilde{R}^*_{MN}$ inside the unit circle is at most N. For systems having a real impulse response, the decomposition can be written

$$H_K(z) = \tilde{R}^*_{MN}(z) + \lambda_N z^{-\nu} \frac{V_N(z)}{V_N(z^{-1})} \tag{2.9}$$

where $V_N(z)$ is formed from the eigenvector V_N as above.

III. Implementation of the CF Method

Given a finite-length impulse response $h_K(n)$, the CF method consists of the following steps. For simplicity, we assume in this description that $h_K(n)$ is real.

The CF Algorithm

1) Set up the Hankel matrix $\mathbf{H}_{\nu,K}$ of (2.4) and compute its eigenvalue λ_N which is the $N+1$ largest in modulus. One way to find λ_N is to compute the $N+1$ smallest (possibly negative) and the $N+1$ largest eigenvalues of the matrix. This can be accomplished by tridiagonal reduction followed by Sturm sequencing, and routines are provided for this in EISPACK [35, subroutines TRED1 and TRIDIB].

2) Compute the eigenvector V_N corresponding to λ_N. This can be done rapidly by inverse iteration [35, subroutines TINVIT and TRBAK1].

3) Evaluate the frequency-response of the optimal (noncausal) Chebyshev approximation (2.9) at $L \gg M + N + 1$ equally spaced points along the unit circle

$$\tilde{R}^*_{MN}(e^{j\omega_k}) = H_K(e^{j\omega_k}) - \lambda_N e^{-j\nu\omega_k} \frac{V_N(e^{j\omega_k})}{V_N(e^{-j\omega_k})}$$

$$\omega_k = \frac{2\pi k}{L}, \qquad k = 0, 1, \cdots, L-1.$$

It is preferable to choose L equal to a power of 2 to allow the use of the fast Fourier transform (FFT) for this and the next step. Note that since $h_K(n)$ is real, $\tilde{R}^*_{MN}(e^{j\omega_k}) = \overline{\tilde{R}^*_{MN}(e^{-j\omega_k})}$, so that only $L/2 + 1$ values need to be computed.

4) Inverse Fourier transform $\tilde{R}^*_{MN}(e^{j\omega_k})$ to obtain the impulse response of the extended rational Chebyshev approximation

$$\tilde{r}^*_{MN}(n) = \mathrm{FFT}^{-1}\{\tilde{R}^*_{MN}(e^{j\omega_k})\} = \frac{1}{L}\sum_{k=0}^{L} \tilde{R}^*_{MN}(e^{j\omega_k}) e^{j\omega_k n}.$$

The first $L/2$ samples, $n = 0, \cdots, L/2 - 1$, correspond to the causal part.

For $\nu \geqslant 0$ ($M \geqslant N - 1$), we have the following.

5) Window $\tilde{r}^*_{MN}$, selecting the causal part, to obtain the impulse response of the Hankel-norm approximation

$$r^{(\mathrm{CF})}_{MN}(n) = \begin{cases} \tilde{r}^*_{MN}(n), & n = 0, \cdots, L/2 - 1 \\ 0, & n = L/2, \cdots, L-1. \end{cases}$$

6) Convert the nonparametric impulse response $r^{(\mathrm{CF})}_{MN}$ to parametric form $\{a_i, b_j\}$, $i = 1, \cdots, N$, $j = 0, \cdots, M$ by Prony's method [7], [36].

For $\nu < 0$ ($M < N - 1$), we have the following.

5') Window $\tilde{r}^*_{MN}$ as

$$\hat{r}^{(\mathrm{CF})}_{N-1,N}(n) = \begin{cases} \tilde{r}^*_{MN}(n+\nu), & n = 0, \cdots, L/2 - 1 \\ 0, & n = L/2, \cdots, L-1. \end{cases}$$

6') Convert the nonparametric impulse response $\hat{r}^{(\mathrm{CF})}_{N-1,N}$ to parametric form $\{a_i, c_j\}$, $i = 1, \cdots, N, j = 0, \cdots, N-1$ by Prony's method, and set $b_j = c_{j-\nu}, j = 0, \cdots, M$.

Discussion

The CF algorithm is defined on the basis of a prescribed order (M, N), and in step 2) above, an error measure $|\lambda_N|$ associated with this order is revealed. An alternative is to prescribe only the difference between the number of poles and zeros (ν), and then decide on the final order after the eigenvalues of $\mathbf{H}_{\nu,K}$ have been inspected. This alternative can lead to the most cost-effective filter designs. For many desired filters, the sequence $\{|\lambda_k|\}_0^{K-\nu}$ drops sharply in magnitude over some small interval, and values of N in this vicinity give efficient designs in terms of order versus error.

A related consideration is that one should ensure $|\lambda_N| < |\lambda_{N-1}|$, since otherwise a degeneracy will occur in which $\tilde{R}_{MN}^*$ has fewer than N stable poles (in the quotient of (2.8), some poles and zeros coalesce). This problem often comes up when $H_K(z)$ is an even function ($h_K(n) = 0$ for n odd) or is an odd function ($h_K(n) = 0$ for n even). It is easily circumvented by taking (M, N) of the form (*odd*, *even*) if H_K is even, and (*even*, *even*) if H_K is odd. There are also instances in which $\tilde{R}_{MN}^*$ has reduced order due to the extreme elements of the eigenvector being zero ($v_N(0) = v_N(K - \nu) = 0$). For a complete treatment of possible degeneracies, see [21].

In steps 3) and 4), it is necessary that the FFT size L be sufficiently large that time-aliasing is negligible. Due to the sampling of the frequency axis inherent in the FFT, the nonparametric impulse response obtained from $\tilde{R}_{MN}^*(e^{j\omega})$ is really proportional to $\Sigma_{l=-\infty}^{\infty} \tilde{r}_{MN}^*(n + lL)$. Since the poles of $\tilde{R}_{MN}^*$ do not lie on the unit circle, increasing L sufficiently will reduce the error due to time-aliasing to any desired level. If d is the smallest distance from a pole of $\tilde{R}_{MN}^*$ to the unit circle, then we desire $(1 - d)^{L/2} \approx 0$.

Since the pole radii are not known in advance, it is useful to estimate the amount of time-aliasing after the fact by means of the formula

$$\mu_{ta} \triangleq \frac{L}{m+1} \frac{\sum_{n=L/2}^{L/2+m} \tilde{r}_{MN}^{*2}(n)}{\sum_{n=0}^{L-1} \tilde{r}_{MN}^{*2}(n)}$$

where m is a positive integer less than $L/2$. (We use the value $m = L/16$.) This is a normalized ratio of the energy where zero is expected and the total energy. We have $0 \leqslant \mu_{ta} \leqslant 1$. When $\mu_{ta} \approx 0$, the amount of time aliasing is negligible.

In step 6), if the eigenvector is numerically accurate, and if μ_{ta} is small, then $r_{MN}^{(\mathrm{CF})}$ is by construction the impulse response of an N-pole M-zero rational filter [and similarly for $\hat{r}_{N-1,N}^{(\mathrm{CF})}$ in step 6′)]. In this situation, it does not matter very much what norm is minimized in obtaining the parametric form of the filter. For this purpose we have chosen Prony's method [7], [28], [34], [36], in which the A and B coefficients are obtained separately by solving two systems of Toeplitz equations of order N and $M + 1$, respectively. Code for the solution of Toeplitz linear equations can be found in [46].

Note that steps 3)-6) perform the spectral factorization needed to select the causal part of $\tilde{R}_{MN}^*(z)$. This approach can be applied to any spectral factorization problem where the number of poles and zeros of the causal part is known in advance. An alternative method for fast spectral factorization (based on the FFT and properties of the ramp cepstrum) has been proposed by Henrici [24] and was used in [22]; however, Henrici's method suffers from time-aliasing generated by zeros near the unit circle in addition to that due to poles. Our method is only sensitive to poles near the unit circle.

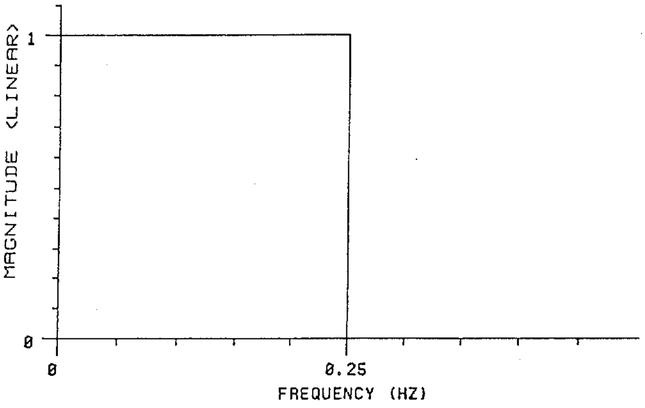

Fig. 1. Ideal low-pass filter magnitude frequency response.

IV. Computed Examples

As has been mentioned earlier, we believe that the CF method may potentially be most competitive in the design of filters with frequency responses that are fairly smooth—at least continuous. Two examples of this can be found in a preliminary version of this report [22]. However, in this section we will consider three relatively standard examples involving discontinuous frequency responses. We do this, first, to show that the CF method is generally applicable, and second, because the familiarity of such applications facilitates comparison with other methods. The design problems selected are

1) minimum-phase recursive low-pass
2) linear-phase recursive low-pass
3) wide-band recursive differentiator.

Example 1: Minimum-Phase Recursive Low-Pass Filter Design

We will use this example to illustrate in detail the various steps of the CF algorithm. In Fig. 1 is shown the ideal low-pass filter magnitude frequency response for a cutoff frequency of one-fourth the sampling rate $f_s \triangleq 1$.

In order to obtain a practical "ideal" minimum-phase impulse response corresponding to Fig. 1, we begin with the function

$$H(\omega) = \begin{cases} 0 \text{ dB}, & 0 \leqslant \omega < \pi/2 \\ -30 \text{ dB}, & \omega = \pi/2 \\ -60 \text{ dB}, & \pi/2 < \omega \leqslant \pi \end{cases}$$

as the desired magnitude frequency response. Thus, we replace the ideal transfer characteristic by one which steps down 60 dB in the frequency domain. This function is then sampled at equally spaced frequencies. For our example, 129 points are used, corresponding to an FFT of length 256. Next, the real-cepstrum method [14], [29] is used to create the minimum-phase complex spectrum exhibiting this magnitude curve. The use of two samples rather than one in the discontinuity serves to reduce time-aliasing. The inverse FFT of the spectrum so obtained yields the initial desired impulse response, and this is shown in Fig. 2(a). The magnitude spectrum of Fig. 2(a) is shown in Fig. 2(b), illustrating the fact that little distortion is incurred at the sample points during the conversion from zero-phase to minimum-phase.

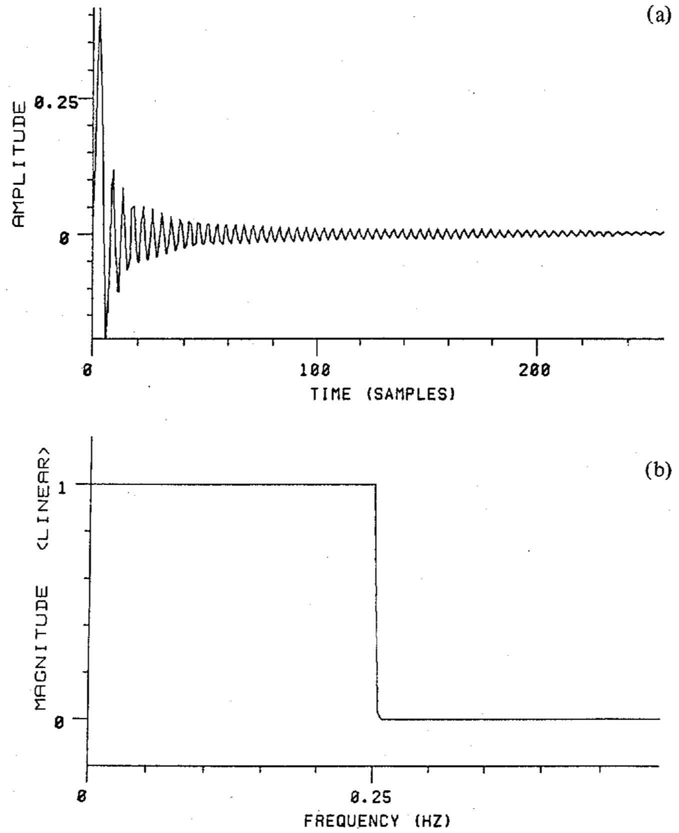

Fig. 2. Minimum-phase ideal low-pass filter obtained by windowing the real cepstrum of the impulse response. (a) Impulse response. (b) Magnitude frequency response.

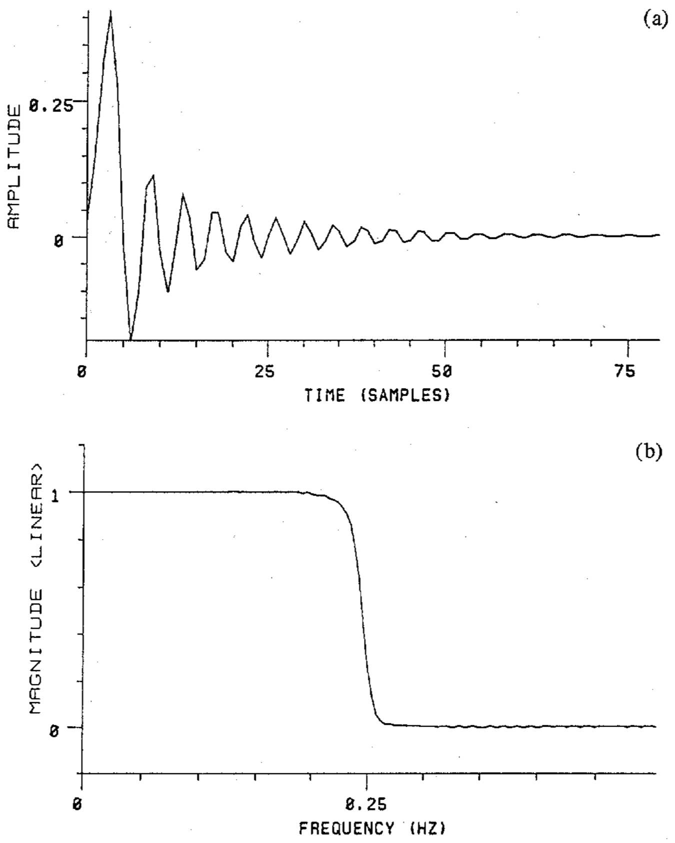

Fig. 3. Hamming-windowed minimum-phase low-pass filter. (a) Impulse response. (b) Magnitude frequency response.

The next step is to window the "ideal" impulse response to the length K desired for use in the CF algorithm. In this case, we choose $K = 79$. The method selected for this windowing consists of multiplying the function of Fig. 2(a) by half of a Hamming window. The resulting impulse response and the corresponding magnitude spectrum are shown in Fig. 3(a) and (b).

We now use the CF method to obtain a 7-pole, 6-zero digital filter which approximates the filter of Fig. 3. First, the 80×80 Hankel matrix is formed, and its 16 extreme eigenvalues are computed. The magnitudes of all 80 eigenvalues are plotted in Fig. 4. The seventh eigenvalue modulus is $|\lambda_7| = 0.019$. This number provides the magnitude of the all-pass error in the optimum noncausal Chebyshev filter, and equals the Hankel norm of the final approximation error. Thus, we expect about two percent error in the magnitude of the passband. The internal FFT size was chosen to be $L = 512$.

Fig. 5(a) shows the magnitude error

$$|H_K(e^{j\omega_k})| - |\tilde{R}^*_{MN}(e^{j\omega_k})|$$

in the optimum extended rational Chebyshev approximation. When the noncausal part of $\tilde{r}^*_{MN}$ is dropped to obtain $r_{MN}^{(\mathrm{CF})}$, the magnitude error becomes that shown in Fig. 5(b). Note how slightly the magnitude error for the optimum Hankel approximation extends past the bounds for the optimum Chebyshev error.

The causal impulse response $r_{MN}^{(\mathrm{CF})}$ of the optimal Hankel approximation is finally converted to a set of recursive filter coefficients, via Prony's method applied to the first 80 samples

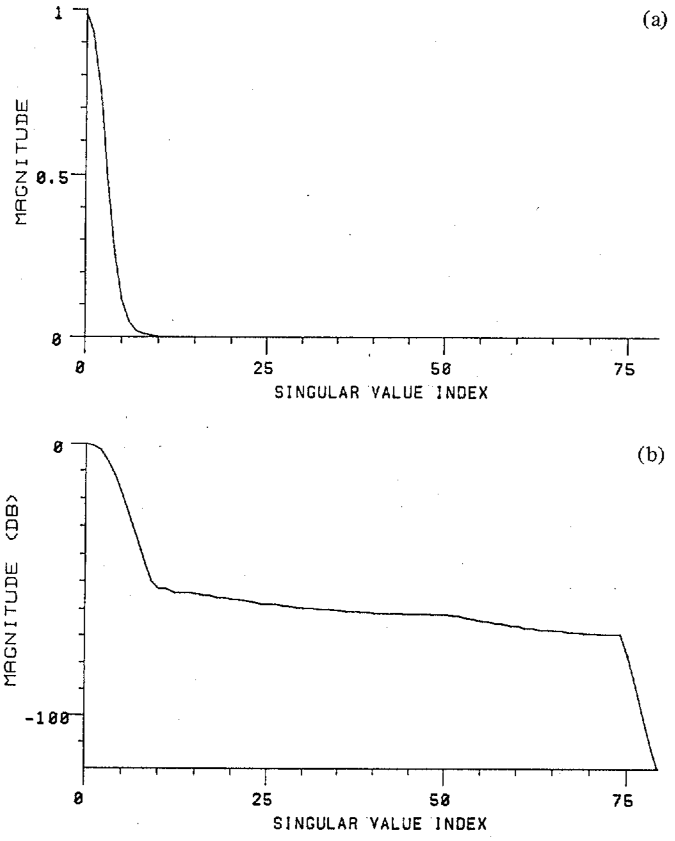

Fig. 4. Singular values of Hankel matrix $\mathbf{H}_{0,80}$ of windowed minimum-phase filter. (a) Linear scale. (b) dB scale.

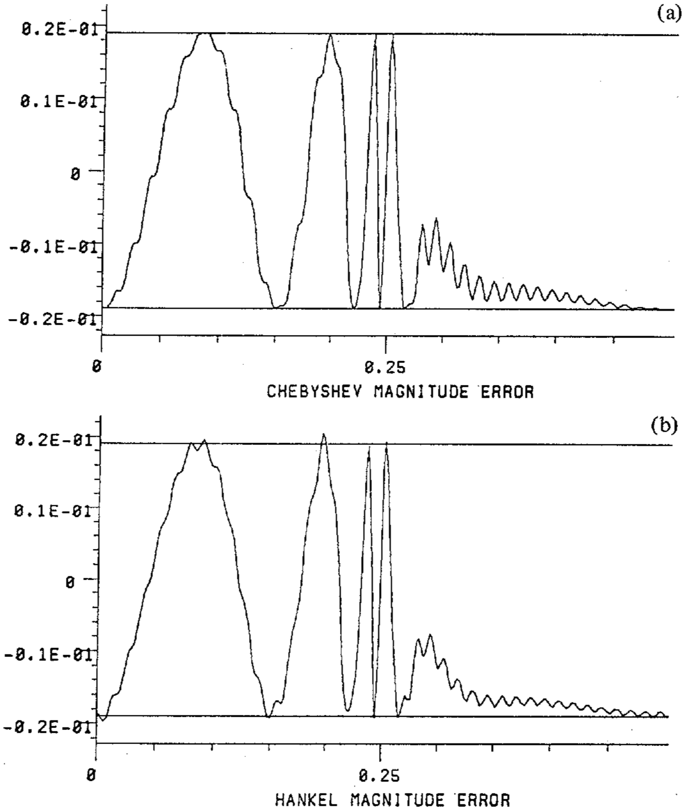

Fig. 5. Magnitude frequency response error. (a) Optimum Chebyshev approximation. (b) CF approximation (optimum Hankel norm).

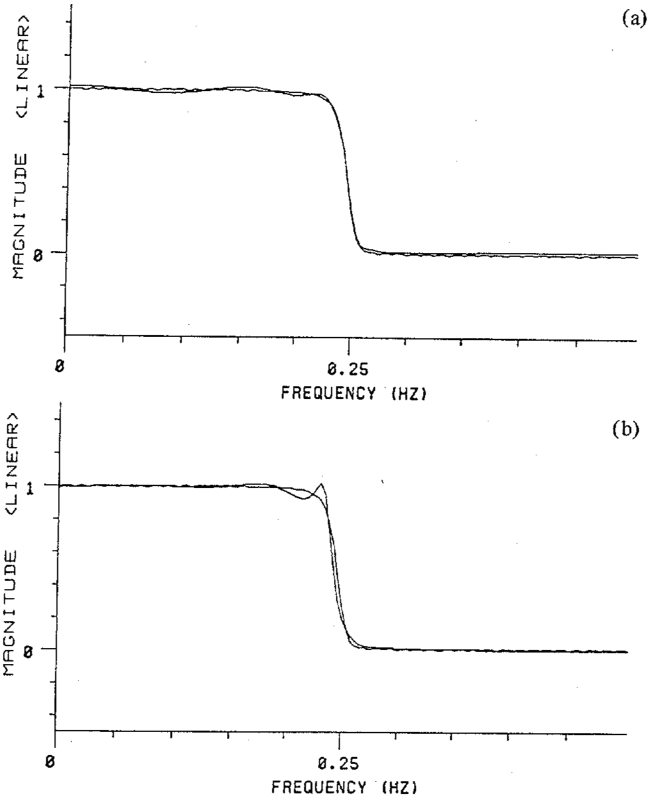

Fig. 6. Magnitude frequency response fit for Example 1. (a) CF. (b) Equation error.

of $r_{MN}^{(CF)}$. The error due to this conversion is $\|r_{MN}^{(CF)} - \hat{r}_{MN}^{(CF)}\|_2 = 0.00012$, where $r_{MN}^{(CF)}$ denotes the impulse response obtained nonparametrically, and $\hat{r}_{MN}^{(CF)}$ denotes the impulse response of the filter computed by Prony's method. (The norm is measured over the first 512 samples of each impulse response.) The good match by Prony's method indicates numerical success of the preceding steps, and that L is sufficiently large.

The final frequency response, overlaid with the desired frequency response, is shown in Fig. 6(a). Notice that the error is nearly equal ripple at about two percent in the passband, as expected.

The filter design obtained using *equation-error* minimization on the same target spectrum $H_K(e^{j\omega})$ as for the CF method is shown in Fig. 6(b). We chose the equation-error method as a standard for comparison because algorithms in this class (such as Prony's method) seem to be the only other way to obtain unique rational approximations which fit both phase and magnitude and which do not suffer from the possibility of convergence to suboptimal solutions [36]. The equation-error algorithm used is a fast version of the one outlined in [45], and it is described in [36]. Note that there is more error near the passband edge with equation-error minimization, due to the presence of poles nearby. [The equation error is defined as $A(e^{j\omega})(R_{MN}(e^{j\omega}) - B(e^{j\omega})/A(e^{j\omega}))$, which gets weighted toward zero near roots of $A(z)$.] On a Foonly F2 computer, in single precision floating point, the equation-error solution required approximately 2.5 s of CPU time, while the CF algorithm took approximately 70 s (with 60 s spent in the tridiagonalization of the 80 X 80 Hankel matrix).

Although the CF method does not attempt to minimize any kind of log-spectral error, it is often the case in filter design that such an error is most appropriate. For completeness we show the CF and equation-error magnitude fits on a dB vertical scale in Fig. 7. On a log vertical scale, the equation-error method may be preferable to the CF method due to better stopband rejection.

Fig. 8 compares the pole-zero plots for the CF and equation-error methods. The large difference between the two plots suggests that use of the equation-error solution as an initial guess for a gradient-descent algorithm, which explicitly minimizes $\|H(e^{j\omega}) - \hat{H}(e^{j\omega})\|$ with respect to pole positions, may not be effective in general.

Example 2: Linear-Phase Recursive Low-Pass Filter Design

In this example, the goal is to design a *linear phase* recursive low-pass filter. Since the CF method requires a finite impulse response as a starting point, it is good to have an initial target impulse response which is optimal in some sense. The Parks-McClellan-Rabiner (PMR) algorithm [30], [33] provides optimal FIR filters in the sense that the Chebyshev norm of the spectral magnitude error is minimized over filters with exactly linear phase. Since the CF method takes an FIR filter into an IIR filter, preserving the spectrum in a nearly optimal Chebyshev sense, the PMR algorithm provides a good initial condition for this problem. Furthermore, our experience indicates that the amount of computational effort in the two methods is comparable, with the CF algorithm being somewhat more expensive. Thus the PMR algorithm is a well-matched supplement to the CF algorithm.

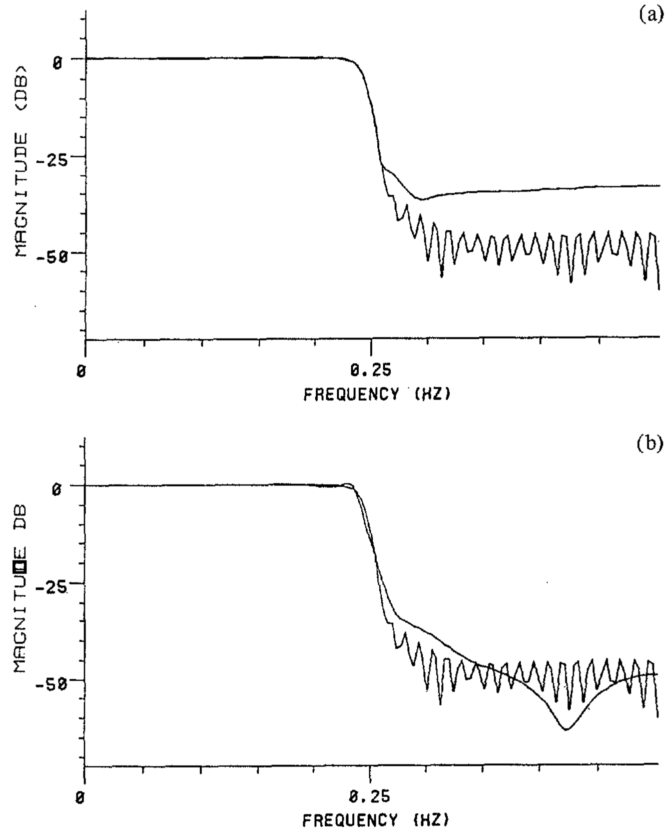

Fig. 7. Magnitude frequency response fit for Example 1 (dB scale). (a) CF. (b) Equation error.

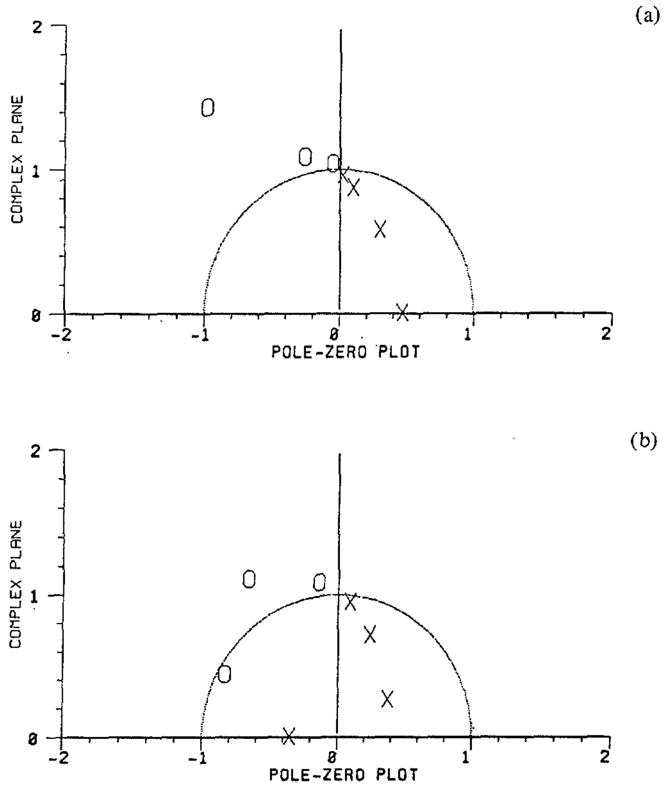

Fig. 8. Poles and zeros for Example 1 (X = pole, O = zero). (a) CF. (b) Equation error.

We begin with an optimal FIR low-pass filter of length $K = 21$. The passband ranges from $f = 0$ to one-tenth the sampling rate $f = f_s/10$, and the stopband is defined from $f = f_s/5$ to $f = f_s/2$. The singular values of the Hankel matrix for this problem are plotted in Fig. 9. In Fig. 10, a comparison between the CF method and the equation-error method is given for the case of a 7-pole, 6-zero approximation to the optimum order 20 FIR filter. The FFT size used is $L = 256$. Fig. 11 gives the same comparison on a dB vertical scale. The impulse response fits for the two methods are shown in Fig. 12, and pole-zero diagrams are shown in Fig. 13. In this example, the CF method clearly outperforms the equation-error method.

Example 3: Wide-Band Differentiator Design

The ideal differentiator has the frequency response

$$H(e^{j\omega}) = j\omega, \qquad -\pi < \omega \leqslant \pi.$$

The design of recursive approximations to $H(e^{j\omega})$ has been addressed by Rabiner and Steiglitz [32], [37], where a method for computing recursive approximations minimizing $\| |H| - |R_{MN}| \|_2$ is proposed. The CF method obtains slightly better approximations of type (2, 2) and (4, 4) with less computational effort. (In the following computations, we used the Henrici spectral factorization and the parametric truncation of $\tilde{R}_{MN}^*$ mentioned above; see [22] for details. But these variations have small effect for this problem.)

We begin by dividing the magnitude characteristic by

$$|H_0(e^{j\omega})| \triangleq |e^{j\omega} - 1|.$$

The resulting quotient is nonzero, so its logarithm exists and we can calculate a minimum-phase transfer function by the real cepstrum method as in Example 1. We approximate this by CF approximations of type (1, 2) and (3, 4) and then multiply these approximations by $e^{j\omega} - 1$. The result should be recursive differentiators of type (2, 2) and (4, 4) that are near optimal in maximum error *weighted* by $|e^{j\omega} - 1|^{-1}$. Since $|e^{j\omega} - 1| \approx |\omega|$ for small ω, the filters should not be far from optimal in the physically more appropriate sense of minimizing the maximum relative error.

The magnitude characteristic in this problem changes slope abruptly at $\omega = \pi$, so H has zero continuous derivatives and the impulse response dies out slowly. We have taken $K = 60$ for the (2, 2) approximation and $K = 120$ for the (4, 4) approximation, and 1024 points in all FFT's. Even these values are not quite enough to make $H - H_K$ negligible. The approximations are still good, however, and the resulting computation times (using double precision on an IBM 370/168) of approximately 0.8 s and 3 s, respectively, suggest that the CF computation is roughly ten times faster than that of Rabiner and Steiglitz [37].

Figs. 14(a) and 15(a) show the errors in amplitude as a function of ω for the two differentiators. Figs. 14(b) and 15(b) show the corresponding excess phase. These curves show roughly the same behavior as those computed by Rabiner and Steiglitz [32]. The zeros for the (2, 2) approximation are at 1.00000 and -0.67570, the poles are at -0.13841 and -0.72021, and the multiplicative constant is 0.36773. These numbers are each within 1 percent of those in [37] except for

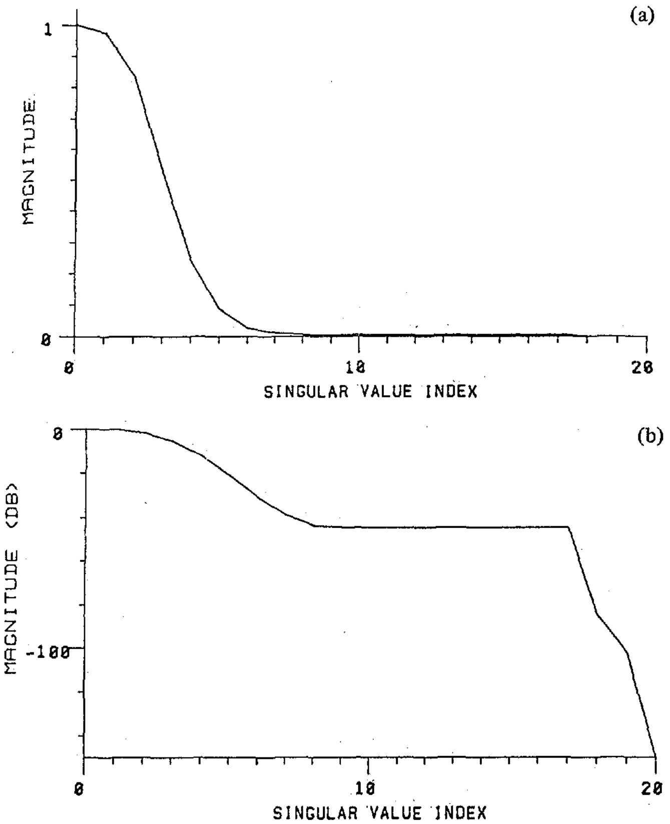

Fig. 9. Singular values of the Hankel matrix $\mathbf{H}_{0,21}$ corresponding to the optimum FIR linear-phase impulse response. (a) Linear scale. (b) dB scale.

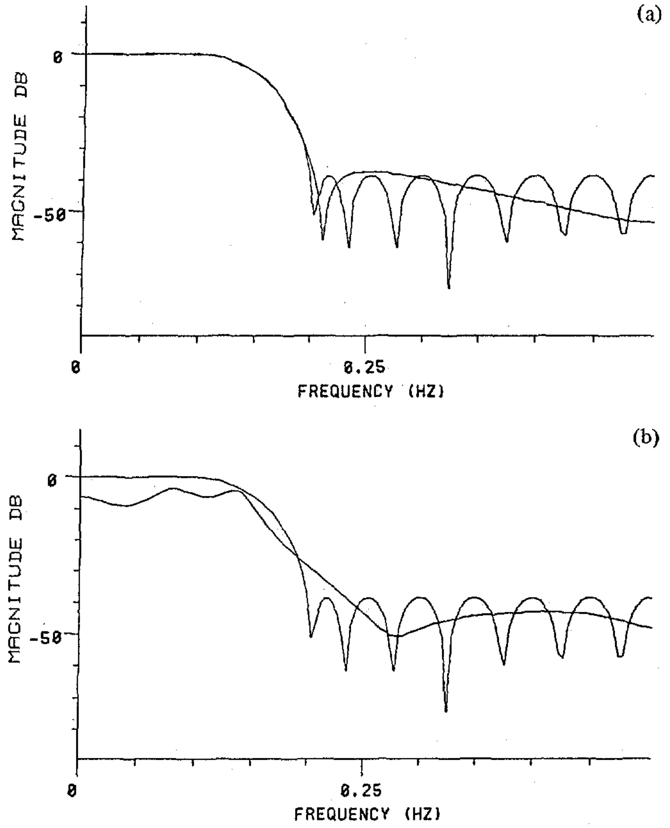

Fig. 11. Magnitude frequency response fit for Example 2 (dB scale). (a) CF. (b) Equation error.

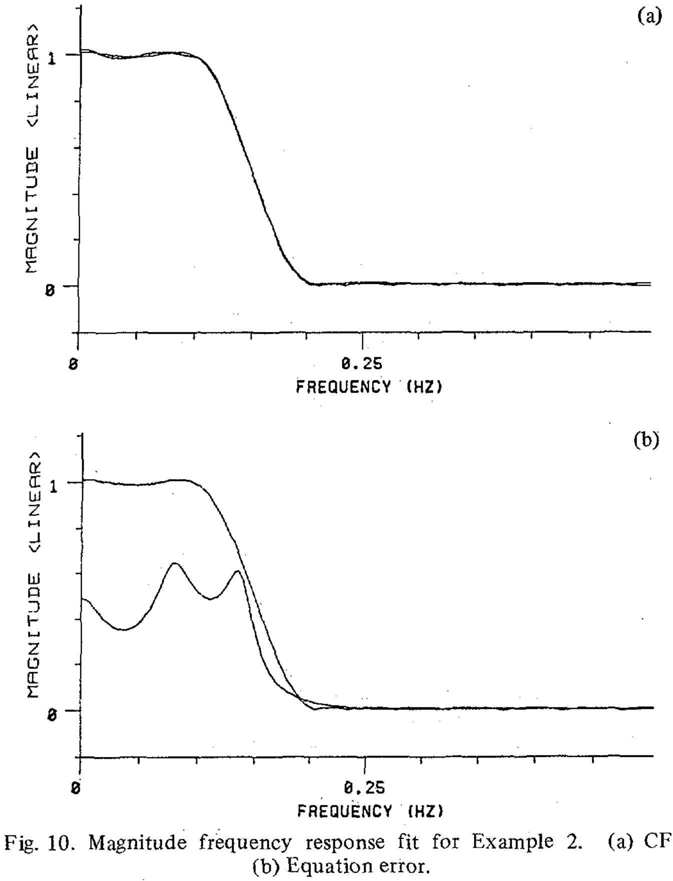

Fig. 10. Magnitude frequency response fit for Example 2. (a) CF. (b) Equation error.

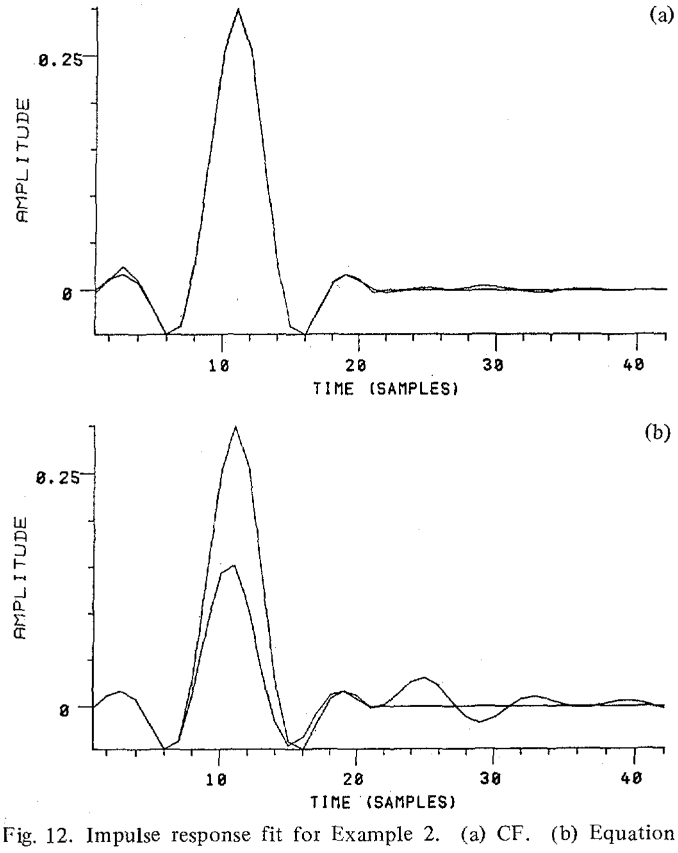

Fig. 12. Impulse response fit for Example 2. (a) CF. (b) Equation error.

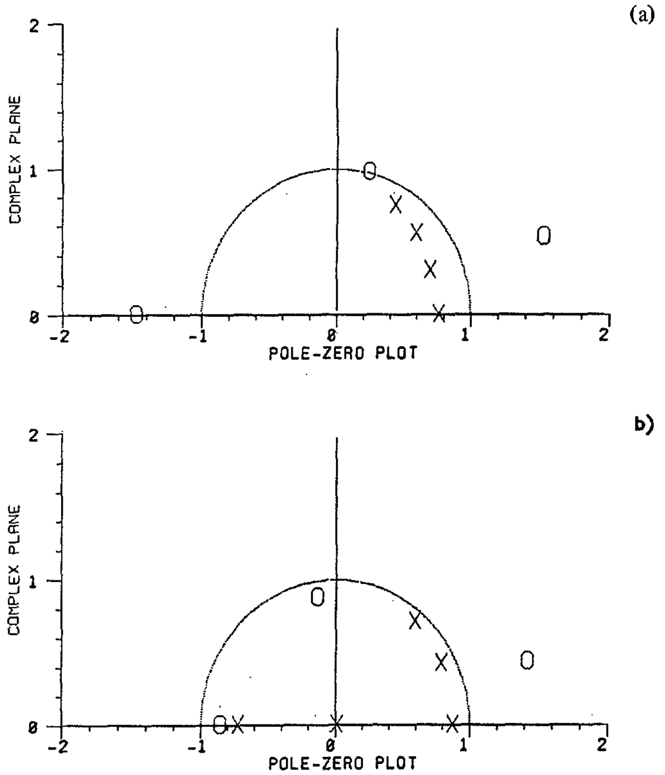

Fig. 13. Poles and zeros for Example 2 (X = pole, O = zero). (a) CF. (b) Equation error.

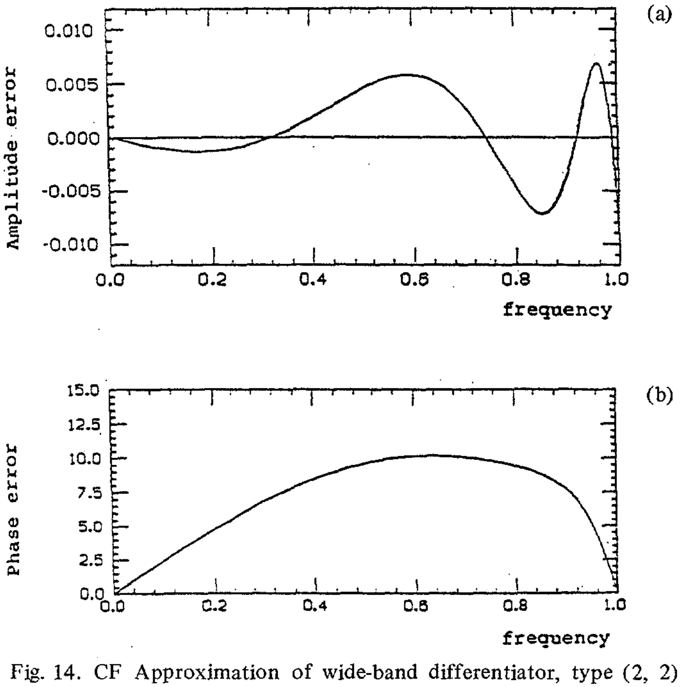

Fig. 14. CF Approximation of wide-band differentiator, type (2, 2). (a) Magnitude error. (b) Phase error.

the pole at −0.13841, which differs by 3 percent. The close agreement is due to the fact that in this example the optimal L_2 approximation, the optimal Chebyshev approximation, and the CF approximation are all roughly the same.

Table I compares the maximum errors of the two CF approximations to those of Rabiner and Steiglitz.

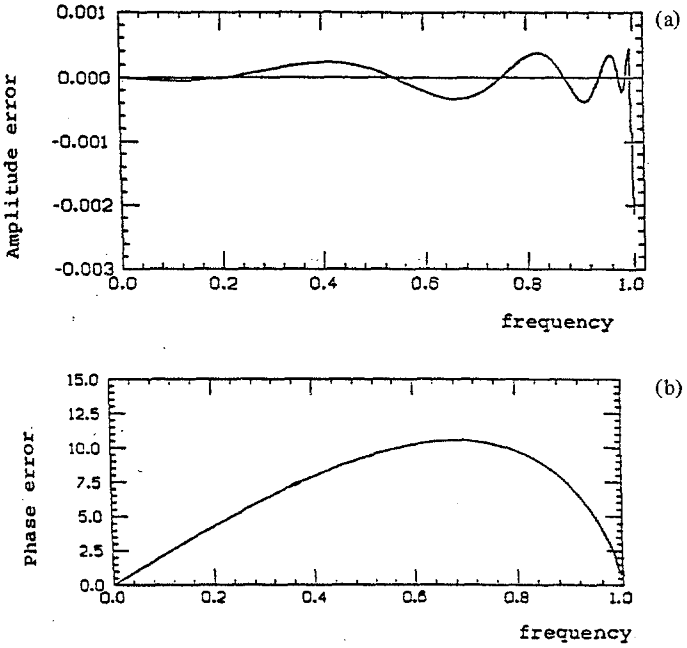

Fig. 15. CF Approximation of wide-band differentiator, type (4, 4). (a) Magnitude error. (b) Phase error.

TABLE I
CF APPROXIMATION TO A WIDE-BAND DIFFERENTIATOR. SEE EXAMPLE 3

(M,N)	K	IBM 370/168 Time	Max. Error	Max. Error Rabiner & Steiglitz
(2, 2)	60	0.8 s	0.011	0.011
(4, 4)	120	3.2 s	0.0021	0.0063

ACKNOWLEDGMENT

We are grateful to Dr. P. van Dooren for his ongoing interest in this project and for bringing many important references to our attention.

REFERENCES

[1] V. M. Adamjan, D. Z. Arov, and M. G. Krein, "Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem," *Math. USSR Sbornik*, vol. 15, pp. 31–73, 1971.

[2] N. Akhieser, *Theory of Approximation.* New York: Frederick Ungar, 1956.

[3] S. Alliney and F. Sgallari, "Chebyshev approximation of recursive digital filters," *Signal Processing*, vol. 2, pp. 317–321, 1980.

[4] J. A. Athanassopoulos and A. D. Waren, "Design of discrete-time systems by mathematical programming," in *Proc. 1968 Hawaii Int. Conf. Syst. Sci.* Honolulu, HI: Univ. Hawaii Press, 1968, pp. 224–227.

[5] L. Barrodale, M. J. D. Powell, and F. D. K. Roberts, "The differential correction algorithm for rational L^∞-approximation," *SIAM J. Numer. Anal.*, vol. 9, pp. 493–504, Sept. 1972.

[6] F. Brophy and A. C. Salazar, "Recursive digital filter synthesis in the time domain," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 45–55, 1974.

[7] C. S. Burrus and T. W. Parks, "Time domain design of recursive digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 137–141, June 1970.

[8] J. A. Cadzow, "High performance spectral estimation–A new ARMA method," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 524–529, Oct. 1980.

[9] C. Carathéodory and L. Fejér, "Über den Zusammenhang der

Extremen von harmonischen Funktionen mit ihrer Koeffizienten und über den Picard-Landauschen Satz.," *Rend. Circ. Mat. Palermo*, vol. 32, pp. 218–239, 1911.

[10] C. K. Chui and A. K. Chan, "A two-sided rational approximation method for recursive digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 141–145, 1979.

[11] D. Clark, "Hankel forms, Toeplitz forms and meromorphic functions," *Trans. Amer. Math. Soc.*, vol. 134, pp. 109–116, 1968.

[12] A. G. Deczky, "Synthesis of recursive digital filters using the minimum *p*-error criterion," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 257–263, 1972.

[13] —, "Equiripple and minimax (Chebyshev) approximations for recursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 98–111, 1974.

[14] Digital Signal Processing Committee, Ed., *Programs for Digital Signal Processing*. New York: IEEE Press, 1979.

[15] D. E. Dudgeon, "Recursive filter design using differential correction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 443–448, 1974.

[16] S. Ellacott and J. Williams, "Rational Chebyshev approximation in the complex plane," *SIAM J. Numer. Anal.*, vol. 13, pp. 310–323, June 1976.

[17] Y. Genin and S. Kung, "A two-variable approach to the model reduction problem with Hankel norm criterion," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 912–924, Sept. 1981.

[18] Y. Genin, "An introduction to the model reduction problem with Hankel norm criterion," in *Proc. European Conf. Circuit Theory and Design*, The Hague, The Netherlands, Aug. 1981.

[19] M. Gutknecht, "Ein Abstiegsverfahren für gleichmässige Approximation, mit Anwendungen," *Diss. ETH Zürich*, 1973.

[20] —, "Non-strong uniqueness in real and complex Chebyshev approximation," *J. Approx. Theory*, vol. 23, pp. 204–213, 1978.

[21] —, "Rational Carathéodory-Fejér approximation on a disk, a circle, and an interval," *J. Approx. Theory*, to be published.

[22] M. Gutknecht and L. N. Trefethen, "Recursive digital filter design by the Carathéodory-Fejér method," Dep. Comput. Sci., Stanford Univ., Standford, CA, Numer. Anal. ms. NA-80-01, 1980.

[23] H. D. Helms, "Digital filters with equiripple or minimax responses," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 87–94, 1971.

[24] P. Henrici, "Fast Fourier methods in computational complex analysis," *SIAM Rev.*, vol. 21, pp. 481–527, 1979.

[25] R. Isermann, Ed., "System identification tutorials," in *Preprints 5th IFAC Symp. Identification*, Darmstadt, Germany, Sept. 24–28, 1979, and *Automatica*, vol. 16, pp. 500–574.

[26] S. Kung, "Optimal Hankel-norm model reductions: Scalar systems," in *Proc. Joint Automat. Contr. Conf.*, San Francisco, CA, 1980.

[27] S. Kung and D. W. Lin, "A state-space formulation for optimal Hankel-norm approximation," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 942–946, Aug. 1981.

[28] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[29] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[30] T. W. Parks and J. H. McClellan, "A program for the design of linear phase finite impulse response (FIR) digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 195–199, Aug. 1972.

[31] L. R. Rabiner, N. Y. Graham, and H. D. Helms, "Linear programming design of IIR digital filters with arbitrary magnitude function," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 117–123, 1974.

[32] L. R. Rabiner and K. Steiglitz, "The design of wide-band recursive and nonrecursive digital differentiators," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 204–209, 1970.

[33] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[34] J. L. Shanks, "Recursion filters for digital processing," *Geophysics*, vol. 32, pp. 33–51, Feb. 1967.

[35] B. T. Smith *et al.*, *Matrix Eigensystem Routines–EISPACK Guide* (Lecture Notes in Comput. Sci. vol. 6), 2nd ed. New York: Springer-Verlag, 1976.

[36] J. O. Smith, "Methods for system identification and digital filter design with application to the violin," Ph.D. dissertation, Dep. Elec. Eng., Stanford Univ., Stanford, CA, 1983.

[37] K. Steiglitz, "Computer-aided design of recursive digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 123–129, 1970.

[38] G. W. Stewart, *Introduction to Matrix Computations*. New York: Academic, 1973.

[39] L. M. Silverman and M. Bettayeb, "Optimal approximation of linear systems," in *Proc. Joint Automat. Contr. Conf.*, San Francisco, CA, 1980.

[40] T. Takagi, "On an algebraic problem related to an analytic theorem of Carathéodory and Fejér and on an allied theorem of Landau" and "Remarks on an algebraic problem," *Japan J. Math.*, vol. 1, pp. 83–93, 1924, and vol. 2, pp. 13–17, 1925.

[41] P. Thajchayapong and P. J. W. Rayner, "Recursive digital filter design by linear programming," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 107–112, 1973.

[42] L. N. Trefethen, "Near-circularity of the error curve in complex Chebyshev approximation," *J. Approx. Theory*, vol. 31, pp. 344–367, 1981.

[43] —, "Rational Chebyshev approximation on the unit disk," *Numer. Math.*, vol. 37, pp. 297–320, 1981.

[44] L. N. Trefethen and M. H. Gutknecht, "The Carathéodory-Fejér method for real rational approximation," *SIAM J. Numer. Anal.*, vol. 20, pp. 420–436, Apr. 1983.

[45] B. Widrow, P. F. Titchener, and R. P. Gooch, "Adaptive design of digital filters," in *Proc. IEEE Conf. Acoust., Speech, Signal Processing*, 1981, pp. 243–246.

[46] S. Zohar, "Fortran subroutines for solution of Toeplitz sets of linear equations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 656–658, Dec. 1979; see also vol. ASSP-28, p. 601, 1980, and vol. ASSP-29, p. 1212, 1981.

[47] M. H. Gutknecht and L. N. Trefethen, "Real and complex Chebyshev approximation on the unit disk and interval," *Bull. Amer. Math. Soc.*, vol. 8, pp. 455–458, May 1983.

[48] —, "Nonuniqueness of best rational Chebyshev approximations on the unit disk," *J. Approx. Theory*, to be published.

Martin H. Gutknecht was born in Berne, Switzerland, on October 1, 1944. He received the diploma in mathematics in 1969, the Doctoral degree in mathematical sciences in 1973, and the *venia legendi* (habilitation) in 1980 from the Swiss Institute of Technology (ETH), Zürich, Switzerland.

Since 1973, he has been a Research Associate and Lecturer at ETH Zürich. In 1976 he was a Postdoctoral Fellow at the Department of Computer Science, University of British Columbia, Vancouver, Canada, and in 1979–1980 he was visiting the Department of Computer Science, Stanford University, Stanford, CA, on a Swiss Senior Research Fellowship. His research interests are in numerical analysis, approximation theory, and numerical methods in engineering.

Julius O. Smith (M'76) was born in Memphis, TN, on March 6, 1953. He received the B.S.E.E. degree from Rice University, Houston, TX, in 1975, and the M.S. and Ph.D. degrees from Stanford University, Stanford, CA, in 1978 and 1983, respectively, on a Hertz Foundation graduate fellowship.

From 1975 to 1977 he worked in the Signal Processing Department at ESL, Sunnyvale, CA, on systems for digital communications. In 1982, he joined the Adaptive Systems Department of Systems Control Technology, Palo Alto, CA, where he has been working in the areas of adaptive filtering and spectrum estimation.

Dr. Smith is a member of Tau Beta Pi and Sigma Xi.

Lloyd N. Trefethen was born in Boston, MA, on August 30, 1955. He received the A.B. degree in applied mathematics from Harvard University, Cambridge, MA, in 1977, and the M.S. and Ph.D. degrees in computer science/numerical analysis from Stanford University, Stanford, CA, in 1980 and 1982, respectively.

His research has been supported by an NSF Graduate Fellowship, a Hertz Foundation Fellowship, and an NSF Postdoctoral Fellowship. His particular interests are approximation theory, applied complex analysis, and finite difference methods for partial differential equations. He is currently with the Courant Institute of Mathematical Sciences, New York University, New York, NY.