MATHEMATICS OF COMPUTATION, VOLUME 29, NUMBER 130
APRIL 1975, PAGES 522–539

Restricted Range Approximation and Its Application to Digital Filter Design

By James T. Lewis*

Abstract. The multiple exchange algorithm for restricted range approximation is discussed. Efficient formulas are derived for the numerical implementation of the method. Discretization effects are analyzed mathematically. The method is applied to a certain problem arising in digital filter design.

1. Introduction. Recently, there has been great interest in the electrical engineering problem of designing nonrecursive digital filters having minimax error [4], [6], [9]. The theory of restricted range approximation [11] in which the approximation is constrained to lie between prescribed upper and lower functions has proved very useful. Certain extensions of the Remes-like single-point exchange algorithm [12] for computing the best approximation have been made. In [5], the multiple exchange algorithm was proposed (see also [3]); this was natural because the multiple exchange algorithm converges in fewer iterations than the single-point exchange method.

Section 2 of this paper contains a statement of the restricted range problem, a description of the multiple exchange algorithm, and a proof of the convergence of the algorithm when the approximation is done on a set with a finite number of points. Section 3 contains a detailed analysis of the discretization error which results when the interval(s) of approximation are replaced by a finite point set. In Section 4, efficient formulas for computing the deviation and the approximation at each iteration are developed; these depend on the form of the basis functions used in the approximation. In Section 5, the digital filter design problem is studied and certain natural questions are considered. Section 6 is a discussion of the results of the implementation of the numerical method.

2. Convergence of the Multiple Exchange Algorithm on Finite Point Sets. The minimax approximation problem with restraining curves can be stated as follows:

$$\min_{a_0,\cdots,a_N} \max_{x\in X} \left| f(x) - \sum_{k=0}^{N} a_k h_k(x) \right|, \tag{2.1}$$

Received November 7, 1973.

AMS (MOS) subject classifications (1970). Primary 65D15, 41A30, 41A50.

Key words and phrases. Approximation with restricted range, computation of best approximations, digital filter design.

* This work was partially supported by the Office of Naval Research under Contract N00014-68-A-0215-0006.

Computer time was sponsored by the National Science Foundation under Facilities Grant GJ-419 with the Department of Computer Science and Experimental Statistics of the University of Rhode Island.

$$\text{subject to} \quad l(x) \leqslant \sum_{k=0}^{N} a_k h_k(x) \leqslant u(x) \quad \text{for all } x \text{ in } X. \tag{2.2}$$

The following hypotheses are made:

H1. $h_0, \cdots, h_N$ form a Chebyshev system on the closed, bounded interval $[a, b]$, that is, $h_0, \cdots, h_N$ are continuous and every nontrivial linear combination $\Sigma_{k=0}^{N} a_k h_k(x)$ has at most N zeros in $[a, b]$.

H2. X is a closed subset of $[a, b]$ with more than $N + 1$ points.

H3. f, l, u are given continuous functions on X and $l(x) \leqslant f(x) \leqslant u(x)$ for all x in X.

H4. There exist $a_0, \cdots, a_N$ such that $l(x) < \Sigma_{k=0}^{N} a_k h_k(x) < u(x)$ for all x in X.

Hypotheses H1–H4 guarantee that the problem (2.1), (2.2) has a unique solution; the proof using weaker hypotheses can be found in [11].

The following characterization theorem from [11] is the foundation for the exchange algorithm.

THEOREM 1 [TAYLOR]. *Assume* H1–H4 *and let*

$$p^*(x) = \sum_{k=0}^{N} a_k^* h_k(x)$$

satisfy

$$l(x) \leqslant p^*(x) \leqslant u(x) \quad \textit{for all } x \textit{ in } X.$$

Set

$$E_+ = \left\{x \in X: f(x) - p^*(x) = \max_{x \in X} |f(x) - p^*(x)|\right\},$$

$$E_- = \left\{x \in X: f(x) - p^*(x) = -\max_{x \in X} |f(x) - p^*(x)|\right\},$$

$$C_+ = \{x \in X: p^*(x) = l(x)\},$$

$$C_- = \{x \in X: p^*(x) = u(x)\}.$$

Let

$$\sigma(x) = \begin{cases} +1 & \textit{if } x \in E_+ \cup C_+, \\ -1 & \textit{if } x \in E_- \cup C_-. \end{cases}$$

Then p^* is the solution of the problem (2.1), (2.2) *if and only if there exist $N + 2$ points $t_0 < t_1 < \cdots < t_{N+1}$ of $E_+ \cup E_- \cup C_+ \cup C_-$ such that $\sigma(t_{i+1}) = -\sigma(t_i)$, $i = 0, \cdots, N$.*

p^* is called a best approximation with restricted range, points in $E_+ \cup E_- \cup C_+ \cup C_-$ are called critical points, and points in $E_+ \cup E_-$ are called extremal points.

The intuitive interpretation of the characterization theorem is the following:

a necessary and sufficient condition for p^* to be a solution of problem (2.1), (2.2) is the existence of $N+2$ points of X where the error $|f-p^*|$ reaches its maximum or p^* hits one of the restraining curves l and u; these occurrences must happen with alternating sign specified by the above σ function.

The multiple exchange algorithm can be viewed as an iterative procedure to locate the critical points of the best approximation; a description of the algorithm for the problem with no restraining curves can be found in [2, p. 97]. To start the algorithm, a set of $N+2$ points $t_0^{(0)} < t_1^{(0)} < \cdots < t_{N+1}^{(0)}$ of X is chosen (called a reference set) and the set of $N+2$ linear equations

$$f(t_i^{(0)}) - \sum_{k=0}^{N} a_k h_k(t_i^{(0)}) = (-1)^i d, \qquad i = 0, \cdots, N+1,$$

is solved, yielding $a_0^{(0)}, \cdots, a_N^{(0)}, d^{(0)}$. Assume the reference deviation $d^{(0)}$ is positive (if $d^{(0)} < 0$ the right-hand side of the equations would be changed to $(-1)^{i+1}d$). Let $p^{(0)} \equiv \Sigma_{k=0}^{N} a_k^{(0)} h_k$. The search for a new reference set is carried out as follows. Consider the case that $t_i^{(0)}$ satisfies $f(t_i^{(0)}) - p^{(0)}(t_i^{(0)}) = +d^{(0)}$. Then points x in X near $t_i^{(0)}$ are examined and $\max_x [f(x) - p^{(0)}(x)]$ and $\max_x [l(x) - p^{(0)}(x)]$ are found (the search is stopped when $f - p^{(0)}$ changes sign). If $[f(x) - p^{(0)}(x)] \leqslant d^{(0)}$ and $p^{(0)}(x) \geqslant l(x)$ for all x near $t_i^{(0)}$ no exchange is made for $t_i^{(0)}$. If

$$\max_x [f(x) - p^{(0)}(x)] - d^{(0)} \geqslant \max_x [l(x) - p^{(0)}(x)],$$

then a point where $\max_x [f(x) - p^{(0)}(x)]$ occurs is exchanged for $t_i^{(0)}$ in the reference set. If $\max_x [l(x) - p^{(0)}(x)] > \max_x [f(x) - p^{(0)}(x)] - d^{(0)}$, then a point, say y, where $\max_x [l(x) - p^{(0)}(x)]$ occurs is exchanged for $t_i^{(0)}$ and the corresponding reference equation is changed to $\Sigma_{k=0}^{N} a_k h_k(y) = l(y)$.

If $t_i^{(0)}$ had satisfied $f(t_i^{(0)}) - p^{(0)}(t_i) = -d^{(0)}$, then $\max_x [p^{(0)}(x) - f(x)]$ and $\max_x [p^{(0)}(x) - u(x)]$ would have been located and the exchange effected in an analogous fashion. In this way, every reference point $t_i^{(0)}$ is (possibly) exchanged. (If no reference points are changed, the algorithm terminates and Theorem 1 guarantees that the solution has been found.) It is still possible that the point of X where

$$\max\left\{\max_x [|f(x) - p^{(0)}(x)| - d^{(0)}], \max_x [l(x) - p^{(0)}(x)], \max_x [p^{(0)}(x) - u(x)]\right\}$$

occurs has not been introduced into the reference set; this must be done. The set of linear equations involving the new reference set is then solved. The procedure is iterated, yielding a sequence $\{t_0^{(i)}, \cdots, t_{N+1}^{(i)}\}$ of reference sets, reference deviations $d^{(i)}$, and approximations $p^{(i)} = \Sigma_{k=0}^{N} a_k^{(i)} h_k$.

In the numerical implementation of the exchange procedure, it is easier to locate the required extrema if the interval(s) of approximation are replaced by a discrete

subset consisting of a finite number of points. In this case, it is easy to prove that the exchange algorithm converges.

THEOREM 2. *Assume* H1–H4 *and let X consist of a finite number of points. Then the reference deviation* $d^{(i)}$ *is strictly increasing and the multiple exchange algorithm reaches the solution of the problem* (2.1), (2.2) *in a finite number of iterations.*

Proof. Assume that the algorithm does not terminate at the ith iteration. We now assume that, for some i, $d^{(i+1)} \leqslant d^{(i)}$, and we seek a contradiction.

Now, $p^{(i)}(x) - p^{(i+1)}(x) = [f(x) - p^{(i+1)}(x)] - [f(x) - p^{(i)}(x)]$ is alternately $\geqslant 0$ and $\leqslant 0$ on the reference set $\{t_0^{(i+1)}, \cdots, t_{N+1}^{(i+1)}\}$. By [10, p. 61], $p^{(i)} \equiv p^{(i+1)}$, a contradiction to the assumption that the algorithm did not terminate (that an exchange was made). Hence, $d^{(i+1)} > d^{(i)}$ for all i. Since X is finite, there are only a finite number of systems of reference set equations possible. Since $d^{(i)}$ is strictly increasing, no system of reference set equations can be repeated. Hence, the algorithm reaches the solution in a finite number of iterations. This proves the theorem.

It should be noted that the tedious proof of the convergence of the multiple exchange method for the case that X is an interval has been carried out in [3].

3. **Discretization Error Analysis.** In this section, the error which arises from solving a sequence of discrete problems (approximation over finite point sets) instead of the continuous problem (approximation over interval(s)) is studied.

Let $X_m = \{x_0, x_1, \cdots, x_m\}$ be a finite point subset of X with $x_0 < x_1 < \cdots < x_m$. The problem we actually solve computationally is

$$\underset{p \in \operatorname{span}\{h_0, \cdots, h_N\}}{\text{minimize}} \ \max_{x_j \in X_m} |f(x_j) - p(x_j)|, \tag{3.1}$$

$$\text{subject to} \quad l(x_j) \leqslant p(x_j) \leqslant u(x_j) \text{ for all } x_j \in X_m. \tag{3.2}$$

Let us consider a sequence X_m, $m = 1, 2, \cdots$, of discrete subsets of X such that

$$\delta_m = \max_{x \in X} \min_{x_j \in X_m} |x - x_j| \longrightarrow 0 \quad \text{as } m \longrightarrow \infty.$$

One would expect that a solution of the discretized problem (3.1), (3.2) would converge to the solution of the original problem (2.1), (2.2) as $m \longrightarrow \infty$; this convergence is established by the subsequent results of this section. We have one further hypothesis to add to H1–H4.

H5. For $m = 1, 2, \cdots$, $X_m = \{x_0, x_1, \cdots, x_m\}$ is a subset of X with $x_0 < x_1 < \cdots < x_m$ and such that $\delta_m = \max_{x \in X} \min_{x_j \in X_m} |x - x_j|$ tends to 0 as m tends to ∞.

In the following, the norm used is the uniform norm on X, i.e., $\|f\| = \max_{x \in X} |f(x)|$.

LEMMA 1. *Assume* H1–H5. *For each positive integer m, let* p_m *be a solution of problem* (3.1), (3.2). *Then there exists a constant A such that* $\sup_{m \geq 1} \|p_m\| \leq A$.

The proof of this lemma will be omitted since it is similar to the proof of Lemma 2 in [7].

Even though $l(x_j) \leq p_m(x_j) \leq u(x_j)$ for all $x_j \in X_m$, it may happen that p_m does not satisfy the constraints on all of X. We now develop bounds on the constraint violation. Let $\omega(g; \delta)$ denote the modulus of continuity of a function g on X; i.e.,

$$\omega(g; \delta) = \max\{|g(x) - g(y)|: x, y \text{ in } X, |x - y| \leq \delta\}.$$

Denote by $\Omega(\delta)$ the joint modulus of continuity of the Chebyshev system $\{h_0, \cdots, h_N\}$; i.e.,

$$\Omega(\delta) = \max_{0 \leq i \leq N} \omega(h_i; \delta).$$

For a continuous function g on the closed, bounded set X, $\omega(g; \delta) \longrightarrow 0$ as $\delta \longrightarrow 0$. Furthermore, $\Omega(\delta) \longrightarrow 0$ as $\delta \longrightarrow 0$ since the h_i are continuous on X and there are a finite number of them.

THEOREM 3. *Assume* H1–H5 *and let* p_m, $m = 1, 2, \cdots$, *be a solution of problem* (3.1), (3.2). *Then there is a constant C independent of m such that for all x in X*

(i) $u(x) - p_m(x) \geq -C \cdot \Omega(\delta_m) - \omega(u; \delta_m)$,

(ii) $p_m(x) - l(x) \geq -C \cdot \Omega(\delta_m) - \omega(l; \delta_m)$.

Proof. For $x \in X$ let x_j in X_m satisfy $|x - x_j| \leq \delta_m$. Then

$$[p_m - u](x) \leq [p_m - u](x_j) + \omega(p_m - u; \delta_m) \leq \omega(p_m; \delta_m) + \omega(u; \delta_m)$$

since $x_j \in X_m$. Let $p_m(x) = \Sigma_{i=0}^{N} a_{i,m} h_i(x)$. Then for $x, y \in X$ with $|x - y| \leq \delta_m$,

$$|p_m(x) - p_m(y)| \leq \sum_{i=0}^{N} |a_{i,m}| \cdot |h_i(x) - h_i(y)| \leq \Omega(\delta_m) \cdot \sum_{i=0}^{N} |a_{i,m}|.$$

Since $\{p_m\}$ is uniformly bounded by Lemma 1, $\Sigma_{i=0}^{N} |a_{i,m}|$ is uniformly bounded in m. So there exists C such that $\omega(p_m; \delta_m) \leq C \cdot \Omega(\delta_m)$. This completes the proof of (i); (ii) is established similarly.

If we make further assumptions on X, X_m, l, u, and $\{h_0, \cdots, h_N\}$, we can obtain better estimates of the constraint violation.

THEOREM 4. *Assume* H1–H5 *and let* p_m *be a solution of problem* (3.1), (3.2). *Assume that X is the union of a finite number of closed intervals. Assume each* X_m *contains the endpoints of these intervals. Let l, u,* $h_0, \cdots, h_N$ *be twice continuously differentiable on X. Then there exists a constant B independent of m such that for all x in X*

(i) $u(x) - p_m(x) \geqslant -B \cdot (\delta_m)^2$,

(ii) $p_m(x) - l(x) \geqslant -B \cdot (\delta_m)^2$.

Proof. Let x be an interior point of X at which $u - p_m$ attains its minimum (if this occurs at an endpoint of the intervals of X, (i) is clearly true). Let $x_j \in X_m$ be in the interval containing x and satisfy $|x - x_j| \leqslant \delta_m$. Then, by Taylor's Theorem,

$$[p_m - u](x_j) = [p_m - u](x) + [p_m - u]'(x)(x_j - x) + \frac{[p_m - u]''(z)}{2}(x_j - x)^2,$$

where z is between x_j and x. So

$$[p_m - u](x) = [p_m - u](x_j) - \frac{[p_m - u]''(z)}{2}(x_j - x)^2,$$

since $[p_m - u]'(x) = 0$,

$$\leqslant \tfrac{1}{2} \cdot [\|p_m''\| + \|u''\|] \cdot (\delta_m)^2.$$

Since $Lp = p''$ is a continuous linear operator on span $\{h_0, \cdots, h_N\}$, it is bounded, i.e., there exists a constant G such that $\|p''\| = \|Lp\| \leqslant G\|p\|$. Using Lemma 1, we see $\sup_{m \geqslant 1} \|p_m''\| < \infty$ and (i) follows. (ii) is established similarly.

The purpose of the next lemma is to get a polynomial close to p_m which satisfies the constraints on all of X.

LEMMA 2. *Assume* H1–H5, *let* p_m, $m = 1, 2, \cdots$, *be a solution of problem* (3.1), (3.2), *assume that* $u(x) - p_m(x) \geqslant -\epsilon_m$ *and* $p_m(x) - l(x) \geqslant -\epsilon_m$ (*where* $\epsilon_m > 0$) *for all* x *in* X. *Let* $q = \Sigma_{i=0}^{N} a_i h_i$ *be as in* H4 *and set*

$$\gamma = \min\left\{\min_{x \in X}(q(x) - l(x)), \min_{x \in X}(u(x) - q(x))\right\}.$$

Then

$$q_m(x) \equiv p_m(x) + \frac{\epsilon_m}{\gamma + \epsilon_m}[q(x) - p_m(x)]$$

satisfies $l(x) \leqslant q_m(x) \leqslant u(x)$ *for all* x *in* X.

Proof. Note that $\gamma > 0$ by H4 and continuity. Since

$$q_m(x) = \frac{\gamma}{\gamma + \epsilon_m} p_m(x) + \frac{\epsilon_m}{\gamma + \epsilon_m} q(x),$$

q_m is a convex combination of p_m and q and hence its graph lies between that of p_m and q. So, if $l(x) \leqslant p_m(x) \leqslant u(x)$, then we still have $l(x) \leqslant q_m(x) \leqslant u(x)$. Assume $l(x) - \epsilon_m \leqslant p_m(x) < l(x)$ for some $x \in X$. Then $p_m(x) = l(x) - \lambda_m(x)\epsilon_m$ where

$$\lambda_m(x) = \frac{l(x) - p_m(x)}{\epsilon_m} \quad \text{satisfies} \quad 0 < \lambda_m(x) \leqslant 1.$$

Now

$$q_m(x) = l(x) - \lambda_m(x)\epsilon_m + \frac{\epsilon_m}{\gamma + \epsilon_m}\,[q(x) - p_m(x)].$$

So

$$q_m(x) - l(x) = -\lambda_m(x)\epsilon_m + \frac{\epsilon_m}{\gamma + \epsilon_m}\,[(q(x) - l(x)) + (l(x) - p_m(x))]$$

$$\geqslant -\lambda_m(x) \cdot \epsilon_m + \frac{\epsilon_m}{\gamma + \epsilon_m}\,[\gamma + \lambda_m(x) \cdot \epsilon_m] \geqslant 0$$

if and only if

$$\lambda_m(x)\epsilon_m \leqslant \frac{\epsilon_m}{\gamma + \epsilon_m}\,[\gamma + \lambda_m(x) \cdot \epsilon_m] \quad \text{if and only if } \lambda_m(x)\epsilon_m\gamma \leqslant \epsilon_m \cdot \gamma,$$

$$\text{if and only if } \lambda_m(x) \leqslant 1.$$

Similarly, it can be shown that $q_m(x) \leqslant u(x)$ for all $x \in X$.

The following theorem is the main result on the discretization error.

THEOREM 5. *Assume* H1–H5. *Let* p_m, $m = 1, 2, \cdots$, *be a solution of problem* (3.1), (3.2). *Assume that* $u(x) - p_m(x) \geqslant -\epsilon_m$ *and* $p_m(x) - l(x) \geqslant -\epsilon_m$ *for all* $x \in X$, *where* $\epsilon_m > 0$ (*cf. Theorems* 3 *and* 4). *Let*

$$\gamma = \min\left\{\min_{x\in X}(q(x) - l(x)), \min_{x\in X}(u(x) - q(x))\right\},$$

where q *is as in* H4. *Then* $\{p_m\}$ *converges uniformly to* p^*, *the unique solution of problem* (2.1), (2.2), *as* $m \to \infty$, *according to the following estimates*:

(i) $\|f - p^*\| - \|f - p_m\| \leqslant (\epsilon_m/(\gamma + \epsilon_m)) \cdot C_1$,

(ii) $\|f - p_m\| - \|f - p^*\| \leqslant \omega(f; \delta_m) + C_2 \cdot \Omega(\delta_m)$,

(iii) $\|p_m - p^*\| \leqslant C_3 \cdot \epsilon_m/(\gamma + \epsilon_m) + C_4\,\omega(f; \delta_m) + C_5 \cdot \Omega(\delta_m)$.

Here, $C_1, \cdots, C_5$ *are independent of* m *and* $\Omega(\delta)$ *is the joint modulus of continuity of* $\{h_0, \cdots, h_N\}$.

Proof. We first note that for q_m as defined in Lemma 2, $l(x) \leqslant q_m(x) \leqslant u(x)$ for all x in X and

$$\|q_m - p_m\| = \frac{\epsilon_m}{\gamma + \epsilon_m}\,\|q - p_m\| \leqslant \frac{\epsilon_m}{\gamma + \epsilon_m}\,[\|q\| + \|p_m\|]$$

$$\leqslant \frac{\epsilon_m}{\gamma + \epsilon_m} \cdot C_1 \quad \text{from Lemma 1.}$$

Hence, $\|f - p^*\| \leqslant \|f - q_m\| \leqslant \|f - p_m\| + \|p_m - q_m\|$. So

$$\|f - p^*\| - \|f - p_m\| \leqslant \frac{\epsilon_m}{\gamma + \epsilon_m} \cdot C_1$$

and (i) is established. Now, let $x \in X$ be such that $|f - p_m|(x) = \|f - p_m\|$ and $x_j \in X_m$ satisfy $|x - x_j| \leqslant \delta_m$. Then

$$\begin{aligned}\|f - p_m\| &= |f(x) - p_m(x)| \\ &\leqslant |f(x) - f(x_j)| + |f(x_j) - p_m(x_j)| + |p_m(x_j) - p_m(x)| \\ &\leqslant \omega(f; \delta_m) + \max_{x_j \in X_m} |f(x_j) - p_m(x_j)| + \Omega(\delta_m) \cdot C_2 \\ &\leqslant \omega(f; \delta_m) + \max_{x_j \in X_m} |f(x_j) - p^*(x_j)| + \Omega(\delta_m) C_2\end{aligned}$$

by the definition of p_m and the fact that $l(x_j) \leqslant p^*(x_j) \leqslant u(x_j)$ for all $x_j \in X_m$,

$$\leqslant \omega(f; \delta_m) + \|f - p^*\| + \Omega(\delta_m) \cdot C_2,$$

which proves (ii).

In order to establish (iii), we need the strong uniqueness theorem for approximation with restricted range, cf. [11]: There exists a constant $\eta > 0$ such that for all $p = \Sigma_{i=0}^{N} a_i h_i$ satisfying $l(x) \leqslant p(x) \leqslant u(x)$ for all x in X,

$$\|p - p^*\| \leqslant (1/\eta)[\|f - p\| - \|f - p^*\|].$$

Now

$$\begin{aligned}\|p_m - p^*\| &\leqslant \|p_m - q_m\| + \|q_m - p^*\| \\ &\leqslant \|p_m - q_m\| + (1/\eta)[\|f - q_m\| - \|f - p^*\|] \\ &\leqslant \|p_m - q_m\| + (1/\eta)[\|f - p_m\| + \|p_m - q_m\| - \|f - p^*\|] \\ &\leqslant \frac{\eta + 1}{\eta} \|p_m - q_m\| + (1/\eta)[\|f - p_m\| - \|f - p^*\|].\end{aligned}$$

(iii) now follows from the above estimate of $p_m - q_m$ and from (ii). Now $\omega(f; \delta_m) \longrightarrow 0$ as $m \longrightarrow \infty$, $\Omega(\delta_m) \longrightarrow 0$ as $m \longrightarrow \infty$. From Theorem 3, we may take $\epsilon_m = C \cdot \Omega(\delta_m) + \omega(u; \delta_m)$ which tends to 0 as $m \longrightarrow \infty$. Hence $\|p_m - p^*\| \longrightarrow 0$ as $m \longrightarrow \infty$, i.e., p_m converges uniformly to p^* as $m \longrightarrow \infty$.

4. Computational Formulas. An essential part of the numerical procedure described in Section 2 is to solve a set of $N + 2$ linear equations for each iteration of the algorithm. In the digital filter problem to be discussed in detail later, $h_k(x) =$

$\cos 2\pi kx$, $k = 0, 1, \cdots, N$, (x in $[0, 0.5]$) is a basis for the set of approximations and, for h_k having this form, special formulas can be derived for solving the reference set equations. The idea is to first calculate the reference deviation d by a certain formula; then $p(x)$ is obtained by interpolation. The analysis is basically an extension of results, known for the problem with no constraints, to the problem with restraining curves. The formulas are very efficient computationally and are useful for the analysis of Section 5.

Let $\{t_0, \cdots, t_{N+1}\}$ be the reference set at a certain iteration (for convenience, the superscripts have been dropped). The linear equations can be written in general

$$f(t_j) - \sum_{k=0}^{N} a_k \cos 2\pi k t_j = d, \qquad j \in I_P,$$

$$f(t_j) - \sum_{k=0}^{N} a_k \cos 2\pi k t_j = -d, \qquad j \in I_M,$$

$$\sum_{k=0}^{N} a_k \cos 2\pi k t_j = l(t_j), \qquad j \in I_L,$$

$$\sum_{k=0}^{N} a_k \cos 2\pi k t_j = u(t_j), \qquad j \in I_U,$$

where I_p, I_M, I_L, I_U is a partition of the set of indices $\{0, \cdots, N+1\}$. Rearranging and using Cramer's Rule, we obtain

$$d = \frac{\begin{vmatrix} 1 & \cos 2\pi t_0 & \cdots & \cos 2\pi N t_0 & r_0 \\ & \vdots & & & \\ 1 & \cos 2\pi t_{N+1} & \cdots & \cos 2\pi N t_{N+1} & r_{N+1} \end{vmatrix}}{\begin{vmatrix} 1 & \cos 2\pi t_0 & \cdots & \cos 2\pi N t_0 & \epsilon_0 \\ & \vdots & & & \\ 1 & \cos 2\pi t_{N+1} & \cdots & \cos 2\pi N t_{N+1} & \epsilon_{N+1} \end{vmatrix}},$$

where

$$r_j = \begin{cases} f(t_j), & j \in I_P, \\ f(t_j), & j \in I_M, \\ l(t_j), & j \in I_L, \\ u(t_j), & j \in I_U, \end{cases} \quad \text{and} \quad \epsilon_j = \begin{cases} 1, & j \in I_P, \\ -1, & j \in I_M, \\ 0, & j \in I_L, \\ 0, & j \in I_U. \end{cases}$$

Expanding the determinants by minors of the last column, we obtain

$$d = \frac{\sum_{j=0}^{N+1} (-1)^j r_j v_j}{\sum_{j=0}^{N+1} (-1)^j \epsilon_j v_j}, \tag{4.1}$$

where

$$v_j = \begin{vmatrix} 1 & \cos 2\pi t_0 & \cdots & \cos 2\pi N t_0 \\ \vdots & & & \\ 1 & \cos 2\pi t_{j-1} & \cdots & \cos 2\pi N t_{j-1} \\ 1 & \cos 2\pi t_{j+1} & \cdots & \cos 2\pi N t_{j+1} \\ \vdots & & & \\ 1 & \cos 2\pi t_{N+1} & \cdots & \cos 2\pi N t_{N+1} \end{vmatrix}$$

$$= \prod_{l>k;\, l\neq j,\, k\neq j} [\cos 2\pi t_l - \cos 2\pi t_k].$$

If we divide numerator and denominator of (4.1) by $v \equiv \Pi_{l>k}[\cos 2\pi t_l - \cos 2\pi t_k]$, we obtain

$$d = \frac{\sum_{j=0}^{N+1} r_j A_j}{\sum_{j=0}^{N+1} \epsilon_j A_j}, \tag{4.2}$$

where

$$A_j \equiv \frac{(-1)^j v_j}{v} = \frac{1}{\prod_{k=0;\, k\neq j}^{N+1} [\cos 2\pi t_k - \cos 2\pi t_j]}, \tag{4.3}$$

The reference deviation d can be calculated using (4.2), (4.3) and the definitions of r_j and ϵ_j. Then the polynomial $p(x)$ satisfying the reference set equations can be obtained by interpolation at $N + 1$ of the $N + 2$ reference points. If we denote by t_m the reference point not used in the interpolation, we have

$$p(x) = \sum_{j=0;\, j\neq m}^{N+1} R_j \prod_{k=0;\, k\neq j,\, k\neq m}^{N+1} \frac{\cos 2\pi x - \cos 2\pi t_k}{\cos 2\pi t_j - \cos 2\pi t_k},$$

where

$$R_j = \begin{cases} f(t_j) - d, & j \in I_P, \\ f(t_j) + d, & j \in I_M, \\ l(t_j), & j \in I_L, \\ u(t_j), & j \in I_U. \end{cases}$$

This can be written in terms of the previously computed A_j's by dividing by

$$1 \equiv \sum_{j=0;j\neq m}^{N+1} 1 \prod_{k=0;k\neq j;k\neq m}^{N+1} \frac{\cos 2\pi x - \cos 2\pi t_k}{\cos 2\pi t_j - \cos 2\pi t_k}$$

and multiplying numerator and denominator by

$$\prod_{k=0;k\neq m}^{N+1} \frac{1}{\cos 2\pi x - \cos 2\pi t_k}.$$

After some algebra, we obtain

$$p(x) = \frac{\sum_{j=0;j\neq m}^{N+1} R_j A_j \dfrac{\cos 2\pi t_m - \cos 2\pi t_j}{\cos 2\pi t x - \cos 2\pi t_j}}{\sum_{j=0;j\neq m}^{N+1} A_j \dfrac{\cos 2\pi t_m - \cos 2\pi t_j}{\cos 2\pi x - \cos 2\pi t_j}}. \tag{4.4}$$

Formula (4.4) is valid for $x \neq t_j, j = 0, \cdots, m-1, m+1, \cdots, N+1$. For $x = t_j$, we simply set $p(t_j) = R_j$, $j = 0, \cdots, m-1, m+1, \cdots, N+1$. The above formulas for d and p are valid for X, the region of approximation, consisting of intervals, a finite point set, or any other closed subset of $[0, 0.5]$. It should be noted that formulas analogous to the above can also be derived if the approximating functions are algebraic polynomials, $h_k(x) = x^k$.

5. Error Trade-Off. We now discuss in detail the prototype problem arising in the design of digital filters:

$$\min_{a_0,\cdots,a_N} \max_{x\in X} \left| f(x) - \sum_{k=0}^{N} a_k \cos 2\pi k x \right|, \tag{5.1}$$

$$\text{subject to } -\epsilon \leqslant \sum_{k=0}^{N} a_k \cos 2\pi k x \leqslant \epsilon \quad \text{for } x_s \leqslant x \leqslant 0.5, \tag{5.2}$$

where $X = [0, x_p] \cup [x_s, 0.5]$ and

$$f(x) = \begin{cases} 1, & 0 \leqslant x \leqslant x_p, \\ 0, & x_s \leqslant x \leqslant 0.5. \end{cases}$$

Here N, x_p, x_s, and ϵ are given parameters (chosen by the filter designer). $[0, x_p]$ is called the passband, $[x_s, 0.5]$ the stopband, and (x_p, x_s) the transition region. (5.1), (5.2) is a digital filter design problem with horizontal lines as restraining curves in the stopband; see Fig. 1. To obtain the effect of no constraints in the passband, we formally set u = large positive constant, l = large negative constant in the passband.

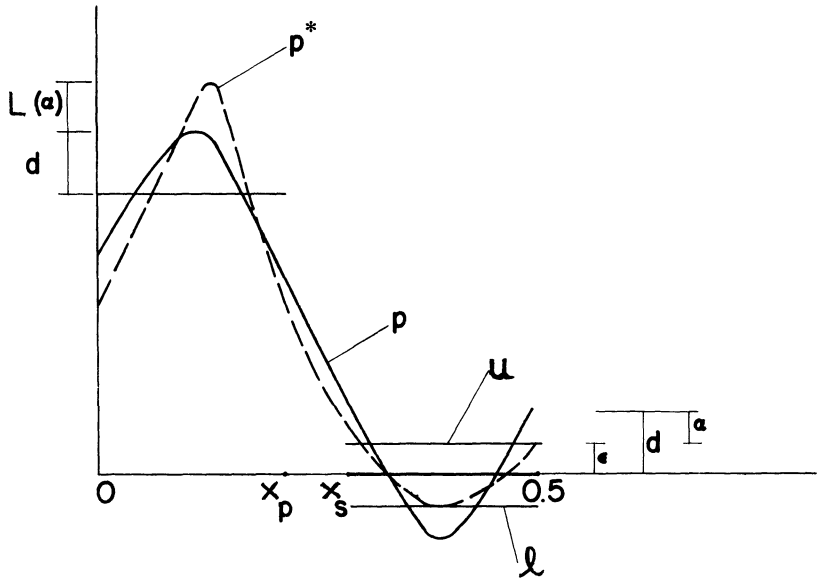

FIGURE 1. *Constrained approximation p^* is shown with unconstrained approximation p. α is the gain and $L(\alpha)$ the loss due to the restraining lines in the stopband*

It is straightforward to verify that hypotheses H1–H4 are satisfied for the problem (5.1), (5.2). The solution of (5.1) is labeled p in Fig. 1. The magnitude of the ripples in the passband is equal to the magnitude of the ripples in the stopband; let d be this deviation. The solution of the constrained problem (5.1), (5.2) is denoted by p^* and is also shown in Fig. 1. ϵ is prescribed and satisfies $0 < \epsilon < d$. $\alpha \equiv d - \epsilon$ is the gain (decrease in stopband ripple) resulting from the constraints. Let d_α be the deviation (passband ripple) resulting from the constrained problem (5.1), (5.2). Then $L(\alpha) \equiv d_\alpha - d$ is the loss (increase in passband ripple) due to the constraints. We now develop some properties of the loss function $L(\alpha)$.

THEOREM 6. *The loss function $L(\alpha)$ described above is an increasing, continuous function for $0 < \alpha < d$. Furthermore, $\lim_{\alpha \to 0} L(\alpha) = 0$ and $\lim_{\alpha \to d} L(\alpha) = 1 - d$.*

Proof. (i) To show L is increasing, let $0 < \alpha_1 < \alpha_2 < d$. If $L(\alpha_2) < L(\alpha_1)$, then $d_{\alpha_2} < d_{\alpha_1}$. Hence, the solution of the problem (5.1), (5.2) with $\epsilon = d - \alpha_2$ has smaller deviation than the solution of the problem (5.1), (5.2) with $\epsilon = d - \alpha_1$, a contradiction. So $L(\alpha_2) \geqslant L(\alpha_1)$, and L is an increasing function.

(ii) Let $0 < \alpha_0 < d$; to show L is continuous at α_0.

Case 1. $\alpha > \alpha_0$. Let p_{α_0} be the solution of the problem (5.1), (5.2) with $\epsilon = d - \alpha_0$. Consider

$$q(x) \equiv \frac{d - \alpha}{d - \alpha_0} p_{\alpha_0}(x).$$

Since q satisfies the constraints of the problem (5.1), (5.2) with $\epsilon = d - \alpha$, we have

$$d_\alpha \leqslant \max_{0 \leqslant x \leqslant x_p} |1 - q(x)| \leqslant 1 - \left[\frac{d-\alpha}{d-\alpha_0}\right] [1 - d_{\alpha_0}]$$
$$= d_{\alpha_0} + \left[\frac{1 - d_{\alpha_0}}{d - \alpha_0}\right](\alpha - \alpha_0).$$

So

$$|L(\alpha) - L(\alpha_0)| = L(\alpha) - L(\alpha_0)$$
$$= d_\alpha - d_{\alpha_0} \leqslant \frac{1 - d_{\alpha_0}}{d - \alpha_0}(\alpha - \alpha_0). \tag{5.3}$$

Case 2. $\alpha < \alpha_0$. Interchanging α_0 and α in the analysis of Case 1 yields

$$|L(\alpha_0) - L(\alpha)| \leqslant \frac{1 - d_\alpha}{d - \alpha}(\alpha_0 - \alpha) \leqslant \left[\frac{1}{d - \alpha_0}\right][\alpha_0 - \alpha]. \tag{5.4}$$

Inequalities (5.3) and (5.4) imply the continuity of $L(\alpha)$ at α_0.

(iii) As in the analysis of Case 1 above, it can be shown that $L(\alpha) = d_\alpha - d \leqslant [(1 - d)/d]\,\alpha$ and $\lim_{\alpha \to 0} L(\alpha) = 0$ follows. It can be shown that if $|p(x)| \leqslant \epsilon$ for all x in $[x_s, 0.5]$, then there exists a constant K such that $|p(x)| \leqslant K\epsilon$ for all x. Hence, as $\alpha \longrightarrow d$ $(\epsilon \longrightarrow 0)$, the solution of the problem (5.1), (5.2) tends to $p(x) \equiv 0$ and so $d_\alpha \longrightarrow 1$. Hence, $\lim_{\alpha \to d} L(\alpha) = 1 - d$. This completes the proof.

From the analysis of Theorem 6, the following rough bound emerges:

$$L(\alpha) \leqslant (1 - d)\alpha/d, \qquad 0 < \alpha < d.$$

In the case that the transition region (x_p, x_s) is symmetric about 0.25, further results can be obtained. A lemma is proved giving information about the unconstrained low pass filter problem (5.1).

LEMMA 3. *Let* $X = [0, x_p] \cup [x_s, 0.5]$ *and assume the transition region* (x_p, x_s) *has* 0.25 *as its midpoint. Then*

(1) *The solution of the unconstrained problem* (5.1) *is of the form*

$$p(x) = 0.5 + \sum_{k\ odd;\, 1 \leqslant k \leqslant N} a_k \cos 2\pi k x.$$

(2) *If* N *is even, there exists a set of* $N + 2$ *extremal points* $\{t_0, \cdots, t_{N+1}\}$ *of the error curve* $f(x) - p(x)$ *associated with the solution of the problem* (5.1) *which are symmetric about* 0.25 *and satisfy* $\sigma(t_j) = -\sigma(t_{j-1})$, $j = 1, \cdots, N + 1$.

(3) *Let* N *be even and* $t_0, \cdots, t_{N+1}$ *the extremal points of* $f(x) - p(x)$ *as in* (2). *If* t_j *and* t_l *are symmetric about* 0.25, *then* $|A_j| = |A_l|$ *where* A_j *and* A_l *are defined by Eq.* (4.3).

Proof. (1) Let

$$p(x) = \sum_{k=0}^{N} a_k \cos 2\pi kx$$

be the solution of the problem (5.1). Then $1 - p(0.5 - x)$ is also a solution of the problem (5.1). By uniqueness,

$$\sum_{k=0}^{N} a_k \cos 2\pi kx = 1 - \sum_{k=0}^{N} a_k \cos 2\pi k[0.5 - x] \tag{5.5}$$

$$= 1 - \sum_{k=0}^{N} (-1)^k a_k \cos 2\pi kx.$$

Hence $a_0 = 0.5$, $a_2 = 0$, $a_4 = 0$, $\cdots$.

(2) From (5.5), we obtain

$$1 - \sum_{k=0}^{N} a_k \cos 2\pi kx = -\left[0 - \sum_{k=0}^{N} a_k \cos 2\pi k[0.5 - x]\right].$$

It is clear from this equation that an error extremal in the passband has a corresponding symmetric (about 0.25) extremal of opposite sign in the stopband and vice versa.

(3) From Eq. (4.3),

$$A_j = 1 \Big/ \prod_{k=0;k\neq j}^{N+1} (\cos 2\pi t_k - \cos 2\pi t_j)$$

and

$$A_l = 1 \Big/ \prod_{k=0;k\neq l}^{N+1} (\cos 2\pi t_k - \cos 2\pi t_l).$$

If t_j and t_l are symmetric about 0.25, since $\cos 2\pi x$ is antisymmetric about 0.25, and since the extremal points are symmetric about 0.25, to each factor $|\cos 2\pi t_{k_1} - \cos 2\pi t_j|$ of $|A_j|$ there corresponds an equal factor $|\cos 2\pi t_{k_2} - \cos 2\pi t_l|$ of $|A_l|$ and vice versa. This completes the proof of the lemma.

THEOREM 7. *Let* $X = [0, x_p] \cup [x_s, 0.5]$ *and assume the transition region* (x_p, x_s) *has* 0.25 *as its midpoint. Then the loss function satisfies* $L(\alpha) \geqslant \alpha$ *for* $0 < \alpha < d$.

Proof.

Case 1. N even. The formula for the reference deviation, cf. Section 4 for the notation and derivation, for the problem with restraining curves l and u is

$$d = \frac{\sum_{i\in I_P \cup I_M} f(t_i)A_i + \sum_{i\in I_L} l(t_i)A_i + \sum_{i\in I_U} u(t_i)A_i}{\sum_{i\in I_P \cup I_M} \epsilon_i A_i},$$

where $\{t_0, \cdots, t_{N+1}\}$ is the reference set described in Lemma 3. If we use the same reference set but new restraining curves $\bar{l}$ and $\bar{u}$, we have

$$\bar{d} = \frac{\sum_{i \in I_P \cup I_M} f(t_i)A_i + \sum_{i \in I_L} \bar{l}(t_i)A_i + \sum_{i \in I_U} \bar{u}(t_i)A_i}{\sum_{i \in I_P \cup I_M} \epsilon_i A_i}.$$

Subtracting, we have

$$\bar{d} - d = \frac{\sum_{i \in I_L} [\bar{l}(t_i) - l(t_i)]A_i + \sum_{i \in I_U} [\bar{u}(t_i) - u(t_i)]A_i}{\sum_{i \in I_P \cup I_M} \epsilon_i A_i}. \tag{5.6}$$

The reference deviation d for the unconstrained problem can be interpreted as the reference deviation for a constrained problem with $l(x) \equiv -d$, $u(x) \equiv d$ in the stopband. If we have a gain of α, then the new restraining curves would be $\bar{l}(x) \equiv -d + \alpha$, $\bar{u}(x) \equiv d - \alpha$. If $\{t_0, \cdots, t_{N+1}\}$ is taken as the initial reference set for the problem with restraining curves $\bar{l}$ and $\bar{u}$, we have from Eq. (5.6)

$$d^{(0)} - d = \alpha \cdot \Big|\sum_{\{i: t_i \in [x_s, 0.5]\}} (-1)^i A_i\Big| \Big/ \Big|\sum_{\{i: t_i \in [0, x_p]\}} (-1)^i A_i\Big|$$

Since the A_i alternate in sign, using Lemma 3, parts 2 and 3, we see $d^{(0)} - d = \alpha$. Since the reference deviation increases with successive iterations

$$L(\alpha) = d_\alpha - d \geqslant d^{(0)} - d = \alpha.$$

Case 2. N odd. Then $N + 1$ is even. Part (1) of Lemma 3 shows that the solution of the unconstrained problem using approximating functions $1, \cdots, \cos 2\pi(N+1)x$ is the same as the solution using $1, \cdots, \cos 2\pi Nx$. Let $d_{\alpha,N+1}$ be the deviation of the solution of the constrained problem using $1, \cdots, \cos 2\pi(N+1)x$ and $d_{\alpha,N}$ the deviation of the solution of the constrained problem using $1, \cdots, \cos 2\pi Nx$. Then, since $d_{\alpha,N+1} \leqslant d_{\alpha,N}$,

$$L(\alpha) = d_{\alpha,N} - d \geqslant d_{\alpha,N+1} - d \geqslant \alpha \quad \text{by Case 1.}$$

This completes the proof.

Theorem 7 lends itself to the interpretation that the imposition of constraints is somewhat unfavorable, since loss is greater than or equal to gain (in the symmetric case). On the other hand, numerical examples have shown that the loss may be less than the gain if the stopband is shorter than the passband; this is intuitively plausible because the constraints are imposed on less than half of the length of X.

The foregoing theory on trade-off error was developed for $X = [0, x_p] \cup [x_s, 0.5]$. Analogous results can be developed for the case (of computational interest)

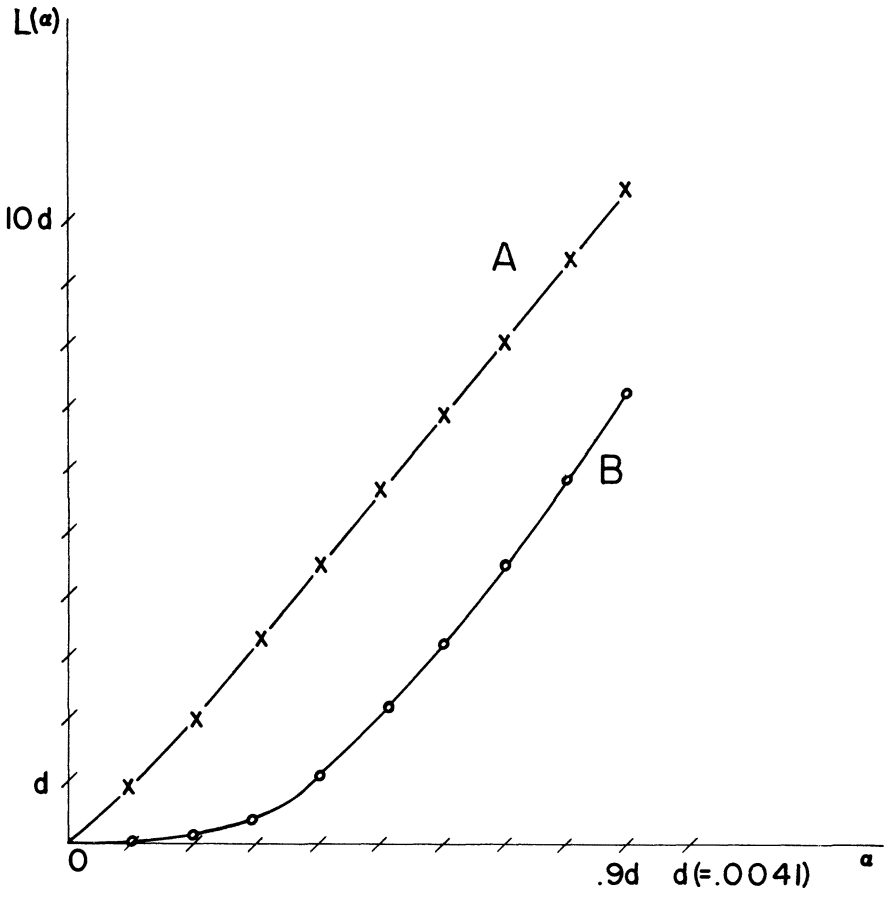

FIGURE 2. *Two plots of* $L(\alpha)$. *Curve* A *results from parameters* $N = 9$, $x_p = 0.0885$, $x_s = 0.2345$. *Curve* B *results from* $N = 9$, $x_p = 0.2655$, $x_s = 0.4115$

that X is a finite point set. To obtain Lemma 3 and Theorem 7, it is necessary to assume that the points of X are symmetric about 0.25.

6. Numerical Implementation and Examples. The numerical procedure for solving the constrained approximation problem, for a finite point set, has been implemented as a double-precision Fortran program and tested numerically on the University of Rhode Island IBM 360/50. The numerical procedure consists of the following steps:

(1) Input of x_p, x_s, N, l, u.

(2) Choice of an initial reference set.

(3) Calculation of d and p at each iteration using formulas (4.2) and (4.4).

(4) Exchange of the reference set at each iteration.

(5) Output of the filter coefficients and deviation (after the procedure has converged).

It should be noted that the program also solves the unconstrained problem

$$\min_{a_0,\cdots,a_N} \max_{x\in X} \left| f(x) - \sum_{k=0}^{N} a_k h_k(x) \right| .$$

This is effected by setting the restraining curve u equal to a large positive constant and l equal to a large negative constant. In solving the unconstrained problem, the initial reference set was chosen as follows. The points in the reference set were equally spaced in the passband and in the stopband. The number to be placed in the passband was determined by using the proportion $x_p/(x_p + (0.5 - x_s))$ of the length of the passband to the length of the approximation region, with the proviso that if fractions arose, the "odd" point was placed in the shorter region. The number of iterations to reach convergence was quite dependent on having the correct number of reference points in each region.

In choosing an initial set for solving the special constrained problem (5.1), (5.2), if α was small, the final reference set (critical points) for the unconstrained problem was found to be an excellent choice. However, for α larger, e.g., $\alpha = 0.9d$ in Fig. 2, it was often found that more points should be placed in the stopband (the constrained region).

x_p = 0.177	x_s = 0.323	x_p = 0.03	x_s = 0.06
N = 9	d = .006482	N = 50	d = .002213
α	L(α)	α	L(α)
.1d = .000648	.000650	.1d = .000221	.000672
.2d = .001296	.001308	.2d = .000443	.001359
.3d = .001945	.001974	.3d = .000664	.002079
.4d = .002593	.002653	.4d = .000885	.002829
.5d = .003241	.003348	.5d = .001106	.003626
.6d = .003889	.004072	.6d = .001328	.004447
.7d = .004537	.004850	.7d = .001549	.005331
.8d = .005186	.005770	.8d = .001770	.006423
.9d = .005834	.033259	.9d = .001991	.028695

x_p = 0.0885	x_s = 0.2345	x_p = 0.2655	x_s = 0.4115
N = 9	d = .004111	N = 9	d = .004111
α	L(α)	α	L(α)
.1d = .000411	.003811	.1d = .000411	.000058
.2d = .000822	.008283	.2d = .000822	.000122
.3d = .001233	.012966	.3d = .001233	.000859
.4d = .001644	.017744	.4d = .001644	.004436
.5d = .002056	.022576	.5d = .002056	.008911
.6d = .002467	.027456	.6d = .002467	.013765
.7d = .002878	.032438	.7d = .002878	.018813
.8d = .003289	.037587	.8d = .003289	.023960
.9d = .003700	.043465	.9d = .003700	.029270

TABLE 1. *Loss function L(α) for four numerical examples*

Plots of the loss function $L(\alpha)$ defined and studied in Section 5 are shown in Fig. 2 for various values of the parameters. A striking feature of the graph B is the abrupt change of slope; this occurred when a critical point moved from the passband to the stopband (constrained region). The slope must become large as the gain α approaches the reference deviation d since $L(\alpha)$ approaches $1 - d$, which is large in comparison with d. Table 1 contains computed values of the loss function $L(\alpha)$ for various values of the parameters. All the numerical examples resulted in $L(\alpha)$ being a convex function.

Acknowledgement. The author is grateful to Donald W. Tufts of the University of Rhode Island Electrical Engineering Department for many valuable discussions.

Department of Mathematics
University of Rhode Island
Kingston, Rhode Island 02881

1. M. C. BUDGE, R. K. CAVIN & D. R. GIMLIN, "Non-recursive filter design via best restricted approximations," Manuscript.

2. E. W. CHENEY, *Introduction to Approximation Theory,* McGraw-Hill, New York, 1966. MR 36 #5568.

3. D. R. GIMLIN, R. K. CAVIN & M. C. BUDGE, "A multiple exchange algorithm for calculation of best restricted approximations," *SIAM J. Numer. Anal.,* v.11, 1974, pp. 219–231.

4. H. D. HELMS, "Digital filters with equiripple or minimax responses," *IEEE Trans. Audio and Electroacoust.,* Vol. AU-19, pp. 87–93, March 1971.

5. H. S. HERSEY, D. W. TUFTS & J. T. LEWIS, "Interactive minimax design of linear-phase nonrecursive digital filters subject to upper and lower function constraints," *IEEE Trans. Audio and Electroacoust.,* pp. 171–173, June 1972.

6. F. K. KUO & J. F. KAISER, *System Analysis by Digital Computer,* Chapter 7, Wiley, New York, 1966.

7. J. T. LEWIS, "Computation of best monotone approximations," *Math. Comp.,* v. 26, 1972, pp. 737–747.

8. T. W. PARKS & J. H. McCLELLAN, "Chebyshev approximation for nonrecursive digital filters with linear phase," *IEEE Trans. Circuit Theory,* vol. CT-19, no. 2, pp. 189–194, March 1972.

9. C. M. RADER & B. GOLD, *Digital Processing of Signals,* McGraw-Hill, New York, 1969.

10. J. R. RICE, *The Approximation of Functions.* Vol. 1: *Linear Theory,* Addison-Wesley, Reading, Mass., 1964. MR 29 #3795.

11. G. D. TAYLOR, "Approximation by functions having restricted ranges. III," *J. Math. Anal. Appl.,* v. 27, 1969, pp. 241–248. MR 41 #2261.

12. G. D. TAYLOR & M. J. WINTER, "Calculation of best restricted approximations," *SIAM J. Numer. Anal.,* v. 7, 1970, pp. 248–255. MR 42 #3978.

13. D. W. TUFTS, J. T. LEWIS & H. S. HERSEY, *Interactive Minimax Design of Nonrecursive Digital Filters,* Report EE 4044/1, Dept. of Electrical Engineering, Univ. of Rhode Island, Kingston, R. I., October 1972.