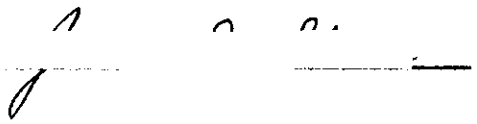
In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institution shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.



A STUDY OF ALGORITHMS FOR OBTAINING BEST UNIFORM APPROXIMATIONS

A THESIS

Presented to

The Faculty of the Graduate Division

by

James Greig Stein

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Applied Mathematics

Georgia Institute of Technology

June, 1963

A STUDY OF ALGORITHMS FOR OBTAINING

BEST UNIFORM APPROXIMATIONS

Date approved by Chairman: May 31, 1963

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to Dr. William J. Kammerer for his time and effort as my thesis adviser. His guidance and encouragement were my greatest assets in writing this thesis. I thank my reading committee, Dr. George C. Caldwell and Dr. David L. Finn. Mrs. Jeanne M. Crawford made most valuable contributions to the final preparation of the manuscript.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Given a function $f(x)$ defined and continuous on a closed and bounded set S contained in E^m and a set of linearly independent functions $f_1, \ldots, f_n$ also continuous on S, it is desired to determine numbers $a_1, \ldots, a_n$ which minimize

$$\max_{x \in S} \left| f(x) - \sum_{i=1}^{n} a_i f_i(x) \right| .$$

This problem, usually referred to as the uniform approximation problem, was first extensively treated by P. L. Tchebycheff [1]. Since the advent of high-speed digital computers, various algorithms for the numerical solution of this problem have appeared in journals. The primary objective of this investigation is to review some of these algorithms and relate and classify them whenever possible.

To facilitate the presentation a brief development of notation as well as some elementary properties of normed linear spaces and convex sets will be given.

Definition 1: A set, L, of elements $f, g, h, \ldots$ is called a linear space over the real field if, and only if, it forms an abelian group with respect to addition, and multiplication is defined between elements of L and scalars $\alpha, \beta, \ldots$, such that

$$1) \quad \alpha(f + g) = \alpha f + \alpha g$$

$$2) \quad (\alpha + \beta)f = \alpha f + \beta f$$

$$3) \quad \alpha(\beta f) = (\alpha\beta)f$$

$$4) \quad 0 \cdot f = 0$$

Definition 2: The linear space L is called a "normed linear space" if there exists a nonnegative function, $\| \ \|$, which satisfies the followings: For every $f, g,$ contained in L and any scalar α

$$\begin{aligned} &1) \quad \| f \| = 0 \quad \text{if, and only if,} \quad f \equiv 0 \\ &2) \quad \| \alpha f \| = |\alpha| \, \| f \| \\ &3) \quad \| f + g \| \leq \| f \| + \| g \| \end{aligned} \tag{I-1}$$

Such a function will be called a norm.

As a consequence of 2) and 3) it follows that

$$\| f - g \| \geq | \| f \| - \| g \| | \tag{I-2}$$

since

$$\| f \| = \| f - g + g \| \leq \| f - g \| + \| g \|$$

and hence

$$\| f - g \| \geq \| f \| - \| g \| .$$

But

$$\| f - g \| = \| g - f \| \geq \| g \| - \| f \|$$

and thus (I-2) is established.

Examples of normed linear spaces.

1) Euclidian n-space. Let E^n be the set of n-tuples of real numbers

$$x = (x_1, \ldots, x_n), \quad y = (y_1, \ldots, y_n), \ldots .$$

Define $x = y$ if, and only if, $x_i = y_i$ for $i = 1,2, \ldots, n$. Define a norm on E^n by

$$\| x \|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} . \tag{I-3}$$

2) The ℓ^p spaces $(1 \leq p \leq \infty)$. Let ℓ^p denote the set of all sequences of real numbers

$$x = \{x_k : k = 1,2, \ldots\}$$

which satisfy

$$\sum_{k=1}^{\infty} |x_k|^p < +\infty .$$

Define a norm on ℓ^p by

$$\| x \|_p = \left(\sum_{k=1}^{\infty} |x_k|^p \right)^{1/p} . \tag{I-4}$$

3) The space C(S). Let $C(S)$ be the set of all real-valued continuous functions defined on a closed and bounded subset S of E_n. Let the norm on $C(S)$ be given by

$$\| f \|_\infty = \max_{t \in S} |f(t)| . \tag{I-5}$$

This norm is commonly referred to as the "uniform norm" (or Tchebycheff norm).

4) The L^p spaces $(1 \leq p \leq \infty)$. Let L^p be the set of all functions which are Lebesque-summable over the interval $[-1,1]$, so that

$$\int_{-1}^{1} |f|^p \, dm < +\infty .$$

Define a norm on $L^p[-1,1]$ by

$$\| f \|_p = \left(\int_{-1}^{1} |f|^p \, dm \right)^{1/p} . \qquad (I\text{-}6)$$

Here it is tacitly assumed that $f = g$ if, and only if, $f(x) = g(x)$ for all x in $[-1,1]$ except perhaps on a set of measure zero.

Definition 3: A vector g is said to be a convex linear combination of the vectors $g_1, g_2, \ldots, g_n$ if there exist scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that $\alpha_i \geq 0$, $i = 1,2, \ldots, n$

$$\sum_{i=1}^{n} \alpha_i = 1$$

and

$$g = \alpha_1 g_1 + \alpha_2 g_2 + \cdots + \alpha_n g_n .$$

A set Σ is called a convex set if, and only if, for $g_1, g_2 \in \Sigma$ any convex linear combination of the form

$$g = \alpha_1 g_1 + \alpha_2 g_2$$

is also contained in Σ.

From a geometric point of view, a convex set in E^n is any subset Σ of E^n for which the line segment joining any two vectors $g_1, g_2, \varepsilon \Sigma$ is also in Σ. Thus the examples Σ_1 and Σ_2 in Figure 1 are convex sets while Σ_3 and Σ_4 are not.

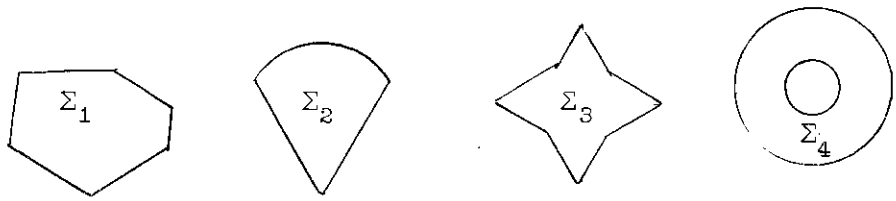


Figure 1. Illustration of Text

A very useful example of a convex set is a closed sphere in a normed linear space, L. Let $f_o \varepsilon L$ and let $\rho > 0$. Then a closed sphere about the point f_o is defined by

$$N(f_o,\rho) = \{f : f \varepsilon L \text{ and } \| f - f_o\| \leq \rho\}.$$

The sphere $N(0,1)$ will be called the unit ball of L.

To see that $N(f_o,\rho)$ is convex, let f,g be any two elements in N. Then the inequality

$$\| \lambda f + (1 - \lambda)g - f_o\| \leq \| \lambda(f - f_o) \| + \| (1 - \lambda) (g - f_o)\|$$

$$= |\lambda| \cdot \| f - f_o\| + |1 - \lambda| \| g - f_o\| \leq 1$$

shows that the line segment $\{\lambda f + (1 - \lambda)g : 0 \leq \lambda \leq 1\}$ is also in N.

Definition 4: A normed linear space is strictly convex if, and only if, when

$$\| f + g \| = \| f \| + \| g \|, \quad \text{for} \quad f,g \neq 0$$

holds then there exists a nonnegative number α such that

$$f = \alpha g .$$

Geometrically the concept of a strictly convex linear space, L, implies that the surface of any closed sphere contained in L cannot contain a line segment. This property is established in the following theorem.

Theorem 1: A linear space, L, is strictly convex if, and only if, the surface of the unit ball, $N(0,1)$, does not contain a line segment.

Proof: We shall prove first that strict convexity implies that the surface of a unit ball cannot contain a line segment. Let us assume that f and g are any two distinct elements on the surface of the unit ball for which the line segment $\{\lambda f + (1 - \lambda)g : 0 \leq \lambda \leq 1\}$ is also on the surface of the unit ball. Thus

$$1 = \| \lambda f + (1 - \lambda)g \| \leq \lambda \| f \| + (1 - \lambda) \| g \| = 1 .$$

In particular, setting $\lambda = 1/2$ we have

$$\| f + g \| = \| f \| + \| g \| .$$

If L were strictly convex, then $f = \alpha g$ for some $\alpha \geq 0$ or $1 = \| f \| = \alpha \| g \|$. This in turn implies $f = g$, which is a contradiction.

To prove the converse we must show that if the surface of the unit ball contains no line segment then L is strictly convex. Let us assume the contrary, that is, that there exist vectors f_1 and f_2 for which

$$\| f_1 + f_2 \| = \| f_1 \| + \| f_2 \|$$

and $f_1 \neq \alpha f_2$ for all $\alpha \geq 0$. We may also assume without loss of generality that $1 = \| f_1 \| \leq \| f_2 \|$. Thus

$$\| f_1 \| + \| f_2 \| = \| f_1 + f_2 \| \leq \| f_1 + \frac{f_2}{\| f_2 \|} \| + (1 - \frac{1}{\| f_2 \|}) \| f_2 \| ,$$

or

$$\| f_1 \| + \| \frac{f_2}{\| f_2 \|} \| \leq \| f_1 + \frac{f_2}{\| f_2 \|} \| \leq \| f_1 \| + \| \frac{f_2}{\| f_2 \|} \| .$$

Letting $f_3 = \frac{f_2}{\| f_2 \|}$ we have

$$\| f_1 + f_3 \| = \| f_1 \| + \| f_3 \|$$

with $\| f_1 \| = \| f_3 \| = 1$ and $f_1 \neq \alpha f_3$ for all $\alpha \geq 0$.

Since the surface of the unit ball contains no line segment we have

$$\| \lambda f_1 + (1 - \lambda) f_3 \| < \lambda \| f_1 \| + (1 - \lambda) \| f_3 \|$$

for $0 < \lambda < 1$. A contradiction is obtained when $\lambda = 1/2$.

Of the examples of normed linear spaces given thus far, the ℓ^p and L^p spaces for $p > 1$ are strictly convex. A proof of this is given by Clarkson [2] by proving they satisfy an even stronger condition, namely, that of being uniformly strictly convex.

The spaces L^1 and $C(S)$ need not be strictly convex. To show this the following examples are given. Consider first the space L^1. Let

$$f(x) = \begin{cases} 1 & \text{if } -1 \leq x < 0 \\ 0 & \text{if } 0 \leq x \leq 1 \end{cases}$$

and define

$$g(x) = 1 - f(x) .$$

It is easy to verify that

$$\| f \|_1 = \| g \|_1 = 1$$

and that the line segment $\{\lambda f + (1 - \lambda)g = 0 \leq \lambda \leq 1\}$ is also contained on the surface of the unit ball. Hence L^1 is not strictly convex.

To show that $C(S)$ is not strictly convex consider $C(S)$ where $S = [0,1]$ and let f and g be given by

$$f(x) = 1$$

and

$$g(x) = x .$$

Then

$$\| f + f \|_\infty = \max_{x \in [0,1]} |1 + x| = 2$$

and

$$\| f \|_\infty = \| g \|_\infty = 1 .$$

Therefore

$$\| f + g \|_\infty = \| f \|_\infty + \| g \|_\infty .$$

But $f \neq \alpha g$ for every $x \in [0,1]$ for any $\alpha \geq 0$; thus $C(S)$ is not strictly convex.

Definition 5: Let M be a subset of a normed linear space, L, which is closed (topologically). If for every $f,g \in M$ and any scalars α, β

$$\alpha f + \beta g \in M$$

then M is called a "subspace" of the space L.

The subspace M is said to be of finite dimension if there exists a set of elements $f_1, f_2, \cdots, f_n$ of M which form a basis for M. Then for every $g \in M$ there exists a unique set of numbers $a_1, a_2, \cdots, a_n$ such that

$$g = a_1 f_1 + a_2 f_2 + \cdots + a_n f_n$$

The set of elements $f_1, f_2, \cdots, f_n$ is said to span the subspace, M.

The fundamental problem of approximation theory can now be considered in the following framework.

Let M be a linear subspace of a normed linear space L and let f be an element of $L - M$. Define the deviation (or distance) of f from M by

$$\rho_M(f) = \min_{g \in M} \| f - g \| .$$

Define the set of best approximations by

$$\Sigma_M(f) = \{g : g \in M \text{ and } \rho_M(f) = \| f - g \|\} .$$

Problem 1: Under what conditions is $\Sigma_M(f)$ nonempty?

Problem 2: Under what conditions will $\Sigma_M(f)$ consist of a single element?

Figure 2 below gives a geometric interpretation of these problems. The quantity $\rho_M(f)$ is the minimal deviation of f from M and this deviation is the same for every element of $\Sigma_M(f)$. While this illustration is of value in interpreting the problem it must be pointed out that the geometry of a normed linear space is not always Euclidean and in fact $\Sigma_M(f)$ can be void.

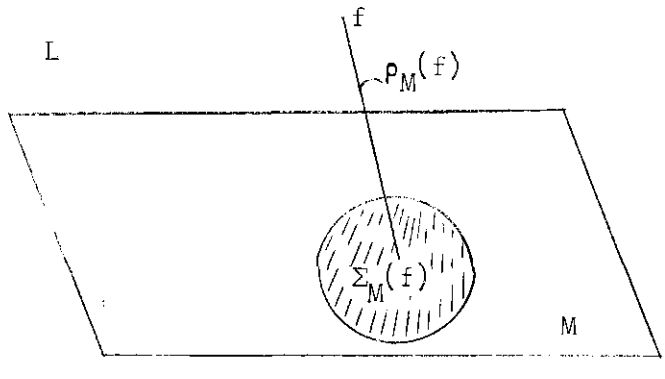


Figure 2

These problems will be considered in the following chapter via existence and uniqueness theorems.

CHAPTER II

EXISTENCE AND UNIQUENESS THEOREMS

The first theorem given establishes the existence of best approximations when M is of finite dimension and is usually referred to as "the fundamental theorem of approximation theory" [3].

Theorem 2: Let L be a normed linear space and let M be a finite dimensional linear subspace of L. Then for every $f \in L$ the set $\Sigma_M(f)$ is nonempty.

Proof: The proof is given in three parts.

1) Define a function φ in E^n by

$$\varphi(a_1, \ldots, a_n) = \| f - g \| = \| f - \sum_{i=1}^{n} a_i f_i \| .$$

This portion of the proof will show that φ is continuous. By definition

$$|\varphi(a_1, \ldots, a_n) - \varphi(b_1, \ldots, b_n)| = | \| f - \sum a_i f_i \| - \| f - \sum b_i f_i \| |$$

and applying (I-2) above

$$| \| f - \sum a_i f_i \| - \| f - \sum b_i f_i \| | \leq \| \sum (b_i - a_i) f_i \|$$

$$\leq \sum |b_i - a_i| \, \| f_i \|$$

$$\leq (\max_i |b_i - a_i|) \sum \| f_i \| .$$

But

$$\sum \| f_i \| = A < \infty .$$

Therefore

$$|\varphi(a_1, \ldots, a_n) - \varphi(b_1, \ldots, b_n)| \leq \max_i |b_i - a_i| A$$

which implies that φ satisfies a Lipschitz condition of order one which in turn implies continuity of φ.

2) Define

$$\psi(\lambda_1, \ldots, \lambda_n) = \| \lambda_1 f_1 + \cdots + \lambda_n f_n \| .$$

Then by Part 1 above ψ is continuous.

Consider ψ on the set K given by

$$K = \{\lambda_1, \ldots, \lambda_n : \sum |\lambda_i|^2 = 1\} .$$

Since K is a closed and bounded set it follows that ψ assumes a minimum on K. Denote this minimum by μ. By definition of the norm, $\mu \geq 0$. Furthermore, since $f_1, \ldots, f_n$ form a basis for M, $\mu \neq 0$. Hence $\mu > 0$.

3) Let $\rho \geq 0$ be the greatest lower bound of $\varphi(a_1, \ldots, a_n)$. This portion of the proof will show that in seeking a minimum it suffices to consider φ defined on a closed and bounded subset of E^n.

From (I-2) above it follows that

$$\varphi(a_1, \ldots, a_n) = \| f - \sum_{i=1}^{n} a_i f_i \| \geq \| a_i f_i + \cdots + a_n f_n \| - \| f \| .$$

Then

$$\varphi(a_1, \ldots, a_n) \geq \left(\sum |a_i|^2\right)^{\frac{1}{2}} \left\| \frac{a_1 f_1}{\sqrt{\Sigma |a_i|^2}} + \cdots + \frac{a_n f_n}{\sqrt{\Sigma |a_i|^2}} \right\| - \| f \|$$

for

$$\left(\sum_{i=1}^{n} |a_i|^2 \right)^{\frac{1}{2}} \neq 0 .$$

Now, let

$$\psi(\lambda_1, \ldots, \lambda_n) = \left\| \frac{a_1 f_1}{\sqrt{\Sigma |a_i|^2}} + \cdots + \frac{a_n f_n}{\sqrt{\Sigma |a_i|^2}} \right\|$$

where

$$\lambda_i = \frac{a_i}{\sqrt{\Sigma |a_i|^2}}, \qquad (i = 1,2, \ldots, n)$$

and hence,

$$\sum_{i=1}^{n} |\lambda_i|^2 = 1 .$$

Therefore by Part 2) $\psi(\lambda_1, \ldots, \lambda_n)$ assumes a minimum, μ, on the unit sphere of E^n. Thus

$$\varphi(a_1, \ldots, a_n) \geq \sqrt{\Sigma |a_i|^2}\, \mu - \| f \| .$$

Recall that $\rho \geq 0$ is by definition the greatest lower bound of $\varphi(a_1, \ldots, a_n)$. For all vectors $(a_1, a_2, \ldots, a_n)$ which satisfy

$$\sqrt{\sum_{i=1}^{n} |a_i|^2} > \frac{1}{\mu} (\rho + 1 + \| f \|$$

we have

$$\varphi(a_1, \ldots, a_n) > \frac{1}{\mu}(\rho + 1 + \| f \|)\mu - \| f \|$$

or

$$\varphi(a_1, \ldots, a_n) > \rho + 1 .$$

Hence it suffices to consider φ only on the closed sphere

$$\{(a_1, a_2, \ldots, a_n) : \sum |a_i|^2 \leq \frac{1}{\mu}(\rho + 1 + \| f \|)\},$$

and on this set φ assumes a minimum. This completes the proof.

Corollary 1: Let $f \in C(S)$ and let M be a finite dimensional linear subspace of $C(S)$. Then the set of best approximations $\Sigma_M(f)$ is non-empty.

Theorem 3: Let M be a finite dimensional linear subspace of a normed linear space, L. If $f \in L - M$ then

1) $\Sigma_M(f)$ is a convex set.

1) $\Sigma_M(f)$ consists of a single point if L is strictly convex.

Proof: 1) Let $N(f, \rho_M(f))$ be a closed sphere in L. Then $\Sigma_M(f)$ satisfies

$$\Sigma_M(f) = N(f, \rho_M(f)) \cap M .$$

Therefore $\Sigma_M(f)$ is convex since $N(f_j, \rho_M(f))$ and M are convex and the intersection of two convex sets is convex.

2) Assume that L is strictly convex and let $f_1, \ldots, f_n$ form a basis for the subspace M. Suppose there exist two vectors g_1, g_2 in

$\Sigma_M(f)$. Then

$$\| f - \sum_{i=1}^{n} a_i f_i \| = \| f - g_1 \| = \rho_M(f)$$

and

$$\| f - \sum_{i=1}^{n} b_i f_i \| = \| f - g_2 \| = \rho_M(f) .$$

Combining these last two equations yields

$$\| f - \sum (\frac{a_i + b_i}{2}) f_i \| \leq \frac{1}{2} \| f - g_1 \| + \frac{1}{2} \| f - g_2 \| = \rho_M(f) .$$

Hence

$$\| f - \sum (\frac{a_i + b_i}{2}) f_i \| = \rho_M(f),$$

for otherwise $\sum (\frac{a_i + b_i}{2}) f_i$ would be a better approximation to f than $\sum a_i f_i$.

Now, since L is strictly convex and

$$\| f - \sum (\frac{a_i + b_i}{2}) f_i \| = \frac{1}{2} \| f - \sum a_i f_i \| + \frac{1}{2} \| f - \sum b_i f_i \|$$

it follows that

$$f - \sum a_i f_i = \alpha \{ f - \sum b_i f_i \} .$$

If $\alpha \neq 1$ then f can be expressed as a linear combination of the f_i contrary to the assumption that $f \in L - M$. This implies $\alpha = 1$ giving in turn

$$\sum (b_i - a_i) f_i = 0$$

or

$$a_i = b_i, \qquad (i = 1,2, \ldots, n) .$$

Thus $\Sigma_M(f)$ consists of a single vector.

In Theorem 3 above it was established that $\Sigma_M(f)$ was nonempty if L was a normed linear space. Now, as a result of Part 2), Theorem 3 we have a sufficient condition that $\Sigma_M(f)$ consists of a single vector g, the condition being that L is strictly convex. Thus in this case a sufficient condition is obtained for uniqueness of a best approximation to f. However, as seen earlier, the class of norms which are strictly convex includes the L^p norms for $1 < p < \infty$, but does not include the uniform norm. But the uniform norm is the one most often used for numerical approximations since it gives an upper bound to the maximum deviation of f from M. Hence, the remainder of this paper will be concerned only with the uniform norm.

In order to have uniqueness for best approximations in the uniform norm further restrictions on M are needed.

Definition 8: Let M be a finite dimensional linear subspace of a normed linear space $C(S)$, so that there exists a set of vectors $f_1, \ldots, f_n$ which form a basis for M. Then the function

$$P(x;\bar{a}) = a_1 f_1(x) + a_2 f_2(x) + \cdots + a_n f_n(x)$$

will be called a "generalized polynomial" or simply a "polynomial." A polynomial in $\Sigma_M(f)$ $C(S)$ will be called a "polynomial of least deviation from f."

Definition 9: A set of functions $f_1, \ldots, f_n$ in $C(S)$ is said to be "unisolvent" on S if, and only if, for every set of n-distinct points of S, the matrix

$$(f_1, f_2, \ldots, f_n) = \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_n(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_n(x_2) \\ \vdots & \vdots & & \vdots \\ f_1(x_n) & f_2(x_n) & \cdots & f_n(x_n) \end{bmatrix} \qquad \text{(II-1)}$$

is nonsingular.

A set of functions which are unisolvent on a set S are sometimes called a "Tchebycheff system" with respect to the set, S. For future use denote the determinant of the matrix given in (II-1) by $D(x_1, x_2, \ldots, x_n)$.

Theorem 4: If M is the span of a unisolvent set of functions $f_1, \ldots, f_n$, then any other basis for M wil also be unisolvent.

Proof: Any two bases $f_1, f_2, \ldots, f_n$ and $g_1, g_2, \ldots, g_n$ of M are related by

$$(f_1, f_2, \ldots, f_n)\, A = (g_1, g_2, \ldots, g_n)$$

where A is a nonsingular $n \times n$ matrix. If $f_1, f_2, \ldots, f_n$ is unisolvent then the identity

$$\begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_n(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_n(x_2) \\ \vdots & \vdots & & \vdots \\ f_1(x_n) & f_2(x_n) & \cdots & f_n(x_n) \end{bmatrix} A = \begin{bmatrix} g_1(x_1) & g_2(x_1) & \cdots & g_n(x_1) \\ g_1(x_2) & g_2(x_2) & \cdots & g_n(x_2) \\ \vdots & \vdots & & \vdots \\ g_1(x_n) & g_2(x_n) & \cdots & g_n(x_n) \end{bmatrix},$$

implies that $g_1, g_2, \ldots, g_n$ is also unisolvent.

Definition 10: If M has a basis which is unisolvent then M will be said to be a unisolvent subspace.

The following result, due to A. Haar [4], gives a necessary and sufficient condition for uniqueness of the polynomial of least deviation from f.

Theorem 5 (Haar's condition): The best approximation in the uniform norm is unique for every $f \in C(S)$ if, and only if, M is a unisolvent subspace.

Proof: The first portion of the proof will show the necessity of Haar's condition. Assume M is not unisolvent. Then there exist n distinct points $x_1, x_2, \ldots, x_n$ in S such that

$$D(x_1, x_2, \ldots, x_n) = 0 .$$

This implies that the row vectors of the matrix given in (II-1) are linearly dependent. Therefore, there exists a nontrivial set of numbers $c_1, c_2, \ldots c_n$ such that

$$c_1 f_j(x_1) + c_2 f_j(x_2) + \cdots + c_n f_j(x_n) = 0$$

for $j = 1, 2, \ldots, n$. Then

$$\sum_j a_j \left[\sum_i c_i f_j(x_i) \right] = 0$$

for any set $a_1, a_2, \ldots, a_n$.

Interchanging the order of summation gives

$$\sum_{i=1}^{n} c_i \left[\sum_{j=1}^{n} a_j \, f_j(x_i) \right] = 0$$

or in other words

$$\sum c_i \, P(x_i;\bar{a}) = 0 \qquad \text{(II-2)}$$

holds for any polynomial $P(x;\bar{a})$.

The assumption $D(x_1,x_2, \ldots, x_n) = 0$ implies the existence of a nontrivial polynomial, $P(x;\bar{b})$, such that $P(x_i;\bar{b}) = 0$ for $i = 1,2, \ldots, n$. Then since $P(x;\bar{b})$ is continuous on S there exists a constant $\lambda \neq 0$ for which

$$|\lambda \, P(x;\bar{b})| < 1 .$$

It is now possible to use $\lambda \, P(x;\bar{b})$ to construct a function $f \in C(S)$ and then show that there exist infinitely many polynomials which are best approximations to f. The construction of f is as follows.

Let $h(x)$ be any continuous function on S satisfying

1) $\max_{x \in S} |h(x)| = 1$

2) $h(x_i) = \text{sign } c_i, \quad (i = 1,2, \ldots, n).$

The function

$$f(x) = h(x) \{ 1 - |\lambda \, P(x;\bar{b})| \} ,$$

will also satisfy properties 1) and 2).

Hence

$$\rho_M(f) \leq \| f(x) - \lambda P(x;\bar{b}) \|_\infty = 1$$

since

$$|f(x_i) - \lambda P(x_i;\bar{b})| = 1, \quad (i = 1,2, \ldots, n)$$

and

$$\begin{aligned} |f(x) - \lambda P(x;\bar{b})| &\leq |f(x)| + |\lambda P(x;\bar{b})| \\ &= |h(x) \{ 1 - |\lambda P(x;\bar{b})| \}| + |\lambda P(x;\bar{b})| \\ &\leq 1 - |\lambda P(x;\bar{b})| + |\lambda P(x;\bar{b})| = 1 . \end{aligned}$$

If on the other hand there exists a polynomial $P(x;\bar{a})$ such that $\rho_M(f) < 1$ then

$$|f(x_i) - P(x;\bar{a})| = |h(x_i) \{1 - |\lambda P(x_i;\bar{b})|\} - P(x_i;\bar{a})| < 1$$

must hold.

But from $P(x_i;\bar{b}) = 0$ and $h(x_i) = \text{sign } c_i$, it follows that

$$|\text{sign } c_i - P(x_i;\bar{a})| < 1 .$$

This implies that

$$\text{sign } c_i = \text{sign } P(x_i;\bar{a})$$

which contradicts (II-2). Therefore, $\rho_M(f) = 1$ and $\lambda P(x;\bar{b})$ is a polynomial of least deviation from f on S.

Next let $|\varepsilon| < 1$ and consider $\varepsilon \lambda P(x;\bar{b})$ as a candidate for a polynomial of least deviation from f on S.

Then

$$|f(x) - \varepsilon\lambda P(x;\bar{b})| \leq |f(x)| + |\varepsilon\lambda P(x;\bar{b})|$$

$$= 1 - |\lambda P(x;\bar{b})| + |\varepsilon| \, |\lambda P(x;\bar{b})|$$

$$= 1 - (1 - |\varepsilon|) \, |\lambda P(x;\bar{b})| \leq 1 .$$

Thus any polynomial of the form

$$\varepsilon\lambda P(x;\bar{b}), \qquad |\varepsilon| < 1$$

is a best approximation to f on S.

2) The proof of the sufficiency of Haar's condition follows. First consider the following lemma which is of interest in itself since it shows that a polynomial of least deviation form f on S assumes its maximum deviation $\rho_M(f)$ on at least n points of S.

Lemma 1: Suppose $f_1, f_2, \ldots, f_n$ satisfy Haar's condition on S and let $P(x;\bar{a})$ be a polynomial of least deviation from $f \in C(S)$. Then

$$\rho_M(f) = |f(x) - P(x;\bar{a})| \qquad \text{(II-3)}$$

holds for at least n points of S.

The proof is by contradiction. Suppose (II-3) holds only for $x_1, x_2, \ldots, x_m$ where $m < n$. Then choose $m - n$ points of S $x_{m+1}, \ldots, x_n$ such that the combined set $x_1, x_2, \ldots, x_n$ constitute n distinct points of S. Then by Haar's condition, the nonhomogeneous system of equations

$$c_1 f_1(x_i) + c_2 f_2(x_i) + \cdots + c_n f_n(x_i) = f(x_i) - P(x_i;\bar{a}) ,$$

$(i = 1,2, \ldots, n)$ has a unique nontrivial solution for $c_1, c_2, \ldots, c_n$.

Let

$$R(x) = f(x) - P(x;\bar{a})$$

and choose a closed neighborhood N_k about each point x_k $(k = 1,2,\ldots,m)$ such that

$$\mu_k = \min_{x \in N_k} |R(x)| > 0$$

and

$$\min_{x \in N_k} |P(x;\bar{c})| \geq \frac{\rho_M(f)}{2} .$$

Let

$$A_k = \max_{x \in N_k} |P(x;\bar{c})| ,$$

$$A = \max_{x \in N} |P(x;\bar{c})| ,$$

and

$$\bar{\rho}(f) = \max_{x \in N} |R(x)|$$

where

$$N = S - \bigcup_{k=1}^{M} N_k .$$

Then

$$\mu = \rho_M(f) - \bar{\rho}(f) > 0 .$$

Now, choose ε such that

$$0 < \varepsilon < \min \left\{ \frac{\mu}{A}, \frac{\mu_1}{A_1}, \ldots, \frac{\mu_m}{A_m} \right\} \text{ and let } \bar{b} = a + \varepsilon \bar{c} .$$

For each $x \in \bigcup_{k=1}^{m} N_k$ the following hold,

$$|f(x) - P(x;\bar{b})| = |f(x) - P(x;\bar{a}) - \varepsilon P(x;\bar{c})|$$

$$= |R(x) - \varepsilon P(x;\bar{c}|$$

$$\leq |R(x)| \left[1 - \frac{\varepsilon P(x;\bar{c})}{R(x)} \right]$$

$$\leq \rho_M(f) \left[1 - \frac{\varepsilon}{2} \right] < \rho_M(f).$$

When $x \in N$

$$|f(x) - P(x;\bar{b}| = |f(x) - P(x;\bar{a}) - \varepsilon P(x;\bar{c})|$$

$$= |R(x) - \varepsilon P(x;\bar{c})|$$

$$\leq |R(x)| + \varepsilon|P(x;\bar{c}|$$

$$\leq \bar{\rho}(f) + \varepsilon A < \rho_M(f) .$$

Therefore,

$$\max_{x \in S} |f(x) - P(x;\bar{b})| < \rho_M(f)$$

which is a contradiction to the hypothesis that $P(x:\bar{a})$ is a polynomial of least deviation from **f** on S.

Now return to the proof of the sufficiency of Haar's condition. Suppose there exist two polynomials $P(x;\bar{a})$, $P(x;\bar{b})$ of least deviation. Then

$$|f(x) - P(x; \frac{\bar{a} + \bar{b}}{2})| \leq \frac{1}{2} |f(x) - P(x;\bar{a})| + \frac{1}{2}|f(x) - P(x;\bar{b})|,$$

which implies that $P(x; \frac{\bar{a}+\bar{b}}{2})$ is also a polynomial of least deviation. Hence, by Lemma 1 there exist n points of S where

$$|f(x_i) - P(x_i; \frac{\bar{a}+\bar{b}}{2})| = \rho_M(f), \qquad (i = 1,2, \ldots, n),$$

or

$$\frac{1}{2}|f(x_i) - P(x_i;\bar{a})| + \frac{1}{2}|f(x_i) - P(x_i;\bar{b})| = \rho_M(f),$$

which implies that

$$f(x_i) - P(x;\bar{a}) = f(x_i) - P(x_i;\bar{b}) = \pm \rho_M(f).$$

But this yields

$$P(x_i; \bar{a} - \bar{b}) = 0, \qquad (i = 1,2, \ldots, n)$$

which contradicts Haar's condition. This completes the proof.

Thus far S has been defined as a compact set in E^n. As a result of Haar's condition it is now important to characterize those compact sets, S, of E^n for which $C(S)$ contains a T-system containing two or more functions.

When S is a finite interval in E^1, say $S = (-\pi,\pi]$ the functions $1, \cos x, \sin x, \ldots, \cos nx, \sin nx$ will form a T-system with respect to S. Likewise the monomials $1, x, x^2, \ldots x^n$ form a T-system on any set S which contains more than n points. Hence, for every compact set $S \subset E'$ there exists a set of functions in $C(S)$ which form a T-system with respect to S.

The following example illustrates the difficulties which arise when $S \subset E^r$ for $r > 1$. The example is an adaptation of one given by R. C. Buck [6].

Let $S \subset E^2$ be defined by

$$S = L_1 \cup L_2 \cup L_3$$

where

$$L_1 = \{(x,y) : 0 < x \leq 1, \quad y = 0\}$$

$$L_2 = \{(x,y) : -1 \leq x < 0, \quad y = 0\}$$

$$L_3 = \{(x,y) : x = 0, \quad 0 \leq y \leq 1\}.$$

Choose a set of n points $u_1, \ldots, u_n$ such that $u_1, u_2 \in L_3$ and $|u_1| > |u_2|$ and the remaining points are contained in L_1 and L_2. Suppose now that u_2 is moved down on L_3 and to the left on L_2 and that u_1 is moved down on L_3 and to the right on L_1. Next return u_1 and u_2 to L_3 so that $|u_2| > |u_1|$. This process amounts to an exchange of two rows of the determinant, $D(u_1, \ldots, u_n)$ without any two points u_i, u_j coinciding and therefore the sign of $D(u_1, \ldots, u_n)$ must change. But since the function D is continuous this implies that there exists a set of points $v_1, \ldots, v_n$ of S such that $D(v_1, \ldots, v_n) = 0$ and $v_i \neq v_j$ for $i \neq j$; hence, Haar's condition is not satisfied on S by any set of continuous functions $f_1, \ldots, f_n$.

From the above example, and the fact that S is a closed set, it follows that the r-dimensional set S cannot contain any interior points for $r \geq 2$.

This discussion will be concluded with the statement of a theorem of Mairhuber [7] which gives a necessary and sufficient condition on S such that $f_1, \ldots, f_n$ form a T-system on S.

Definition 11: A set S_1 is homeomorphic to a set S_2 if there exists a function φ with domain S_1 and range S_2 such that

1) φ is a continuous mapping of S_1 onto S_2.

2) φ^{-1} exists and is continuous on its domain S_2.

Theorem 6: A compact subset S of E^r containing at least n points, $n \geq 2$, may serve as the domain of definition of a set of real-valued continuous functions $f_1, \ldots, f_n$ which satisfy Haar's condition on S if, and only if, S is homeomorphic to a closed subset of a finite interval. If n is even then S may also be the homeomorphic image of the circumference of a circle.

In Lemma 1 it was shown that the error function, $R(x) = f(x) - P(x;\bar{a})$, assumed its maximum value for at least n points of S when $P(x;\bar{a})$ is a polynomial of least deviation from $f \in C(S)$. Another useful property of $P(x;\bar{a})$ is that of oscillation in sign of the error function. Before proving a general theorem on oscillation a related idea is given in the following lemma.

Lemma 2: Let $f_1, \ldots, f_n$ form a T-system with respect to the closed interval $[A,B]$. Then for each set of $n - 1$ distinct points

$$x_1 < x_2 < \cdots < x_{n-1}, \text{ in } [A,B], \text{ the function}$$

$$D(x) = \begin{bmatrix} f_1(x) & f_2(x) & \cdots & f_n(x) \\ f_1(x_1) & f_2(x_1) & \cdots & f_n(x_1) \\ \vdots & \vdots & & \vdots \\ f_1(x_{n-1}) & f_2(x_{n-1}) & \cdots & f_n(x_{n-1}) \end{bmatrix},$$

satisfies

1) $D(x_i) = 0, \quad (i = 1,2, \ldots, n-1)$

2) $D(x)$ changes sign in passing through x_i if x_i is an interior point of $[A,B]$.

Proof: The proof of 1) follows directly from the fact that two rows of the determinant defining $D(x)$ are the same for $x = x_i$ when $i \in \{1,2, \ldots, n-1\}$.

To prove 2) consider $D(x)$ on the open subintervals (A,x_1), (x_1,x_2), $\ldots$, (x_{n-1},B). $D(x) \not\equiv 0$ for $x \in (x_{k-1},x_k)$ by virtue of Haar's condition. Suppose there exists an $i_o \in \{1,2, \ldots, n-1\}$ such that $D(x)$ does not change sign in passing through x_{i_o}. Then, either $D(x) \geq 0$ or $D(x) \leq 0$ for every $x \in (x_{i_o-1}, x_{i_o+1})$ and $x \neq x_{i_o}$.

Assume $D(x) \geq 0$ and let

$$\bar{D}(x) = D(x, x_1, \ldots, x_{i_o-1}, y, x_{i_o+1}, \ldots, x_{n-1})$$

where here also $y \neq x_{i_o}$ but $y \in (x_{i_o-1}, x_{i_o+1})$. It is easy to verify that $\bar{D}(x)$ is also nonnegative for all x in (x_{i_o-1}, x_{i_o+1}). Thus there exists an $\varepsilon > 0$ such that

$$D(x) - \varepsilon \bar{D}(x)$$

has at least n-zeros in $[A,B]$. But this implies a contradiction to Haar's condition. Here, $D(x)$ must change signs in passing through x_{i_o}. A similar result holds for $D(x) \leq 0$.

Theorem 7: Let $f_1, \ldots, f_n$ form a T-system with respect to the interval $[A,B]$ and let $f \in C[A,B]$. Then if $X = \{x_1, \ldots, x_{n+1}\}$ is a set of $n+1$ points of $[A,B]$, the polynomial, $P(x;\bar{a})$, of least deviation from f on X satisfies

1) $|f(x_i) - P(x_i;\bar{a})| = \rho_X(f)$, $(i = 1,2, \ldots, n+1)$, and

2) $\operatorname{sign} R(x_i) = -\operatorname{sign} R(x_{i+1})$, $(i = 1,2, \ldots, n)$.

Proof: The proof which follows is similar to that given by S. Karlin [8].

To prove 1) suppose there exists an $x_{i_o} \in \{x_1, x_2, \ldots, x_{n+1}\}$ such that

$$|f(x_{i_o}) - P(x_{i_o};\bar{a})| < \rho_X(f) \tag{II-4}$$

Then for $j = 1,2, \ldots, n+1$ but $j \neq i_o$ construct a polynomial $D_j(x)$ as Lemma 2, i.e.,

$$D_j(x) = D(x, x_1, \ldots, x_{i_o-1}, x_{i_o+1}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{n+1})$$

and

$$D_j(x_i) = 0 \quad \text{for} \quad i = 1,2, \ldots, n+1,$$

but $i \neq i_o$ and $i \neq j$. Note also that it is possible to construct $D_j(x)$ such that

$$D_j(x_j) > 0.$$

Next let ε_j be chosen so that

$$\text{sign } \varepsilon_j = \text{sign } [f(x_j) - P(x_j;\bar{a})] .$$

Then by (II-4) above it is possible to choose $|\varepsilon_j|$ so small that

$$|f(x_{i_o}) - P(x_{i_o};\bar{a}) - \varepsilon_j D_j(x_{i_o})| < \rho_X(f) .$$

From which it follows that

$$|f(x_j) - P(x_j;\bar{a}) - \varepsilon_j D_j(x_j)| < \rho_X(f)$$

while

$$|f(x_i) - P(x_i;\bar{a}) - \varepsilon_j D_j(x_i)|$$

$$= |f(x_i) - P(x_i;\bar{a})| \leq \rho_X(f) .$$

Therefore

$$|f(x_i) - P(x_i;\bar{a}) - \sum_{j=1}^{n+1} \varepsilon_j D_j(x_i)| < \rho$$

for $(i = 1,2, \ldots, n+1)$.

But this contradicts the assumption that $P(x;\bar{a})$ is a polynomial of least deviation; hence 1) is established.

To prove 2) let x_{i_o}, x_{i_o+1} be two consecutive points of x_1, $x_2, \ldots, x_{n+1}$ and assume that

$$R(x_{i_o}) > 0 \quad \text{and} \quad R(x_{i_o+1}) > 0 .$$

Define

$$D_{i_o}(x) = D(x, x_1, \ldots, x_{i_o-1}, x_{i_o+2}, \ldots, x_{n+1})$$

so that $D_{i_o}(x) > 0$ for $x \in (x_{i_o-1}, x_{i_o+2})$. Then for $\varepsilon > 0$ sufficiently small the polynomial

$$P(x;\bar{a}) - \varepsilon D_{i_o}(x)$$

will yield the same absolute maximum for the error function

$$\bar{R}(x) = f(x) - P(x;\bar{a}) - \varepsilon D_{i_o}(x)$$

since $D_{i_o}(x_i) = 0$ for $i \in \{1,2, \ldots, n+1\}$ but $i \neq i_o$ and $i \neq i_o + 1$.

But

$$\bar{R}(x_{i_o}) < \rho_X(f) \quad \text{and} \quad \bar{R}(x_{i_o+1}) < \rho_X(f)$$

contradicts Part 1) of the theorem. This completes the proof.

This development will be concluded with a generalization of a well-known theorem due to P. L. Tchebysheff.

Theorem 8: Let $f_1, \ldots, f_n$ form a T-system with respect to the interval $[A,B]$ and let $f \in C[A,B]$. Then the polynomial, $P(x;\bar{a})$, of least deviation from f on $[A,B]$ is uniquely characterized by the property that the error function, $R(x) = f(x) - P(x;\bar{a})$, assumes its maximum value, $\rho_{[A,B]}(f) = \rho$, on at least $n + 1$ points of $[A,B]$ and alternates in sign on these $n + 1$ points.

Proof: (Note that the existence and uniqueness of $P(x;\bar{a})$ follows from Theorems 2 and 5). To prove the sufficiency of the condition let

$x_1 < x_2 < \cdots < x_{n+1}$ be $n + 1$ points of $[A,B]$ and let $P(x;\bar{a})$ be a polynomial defined on $[A,B]$ such that

$$|R(x_i)| = \rho, \quad (i = 1,2, \ldots, n+1)$$

and $R(x)$ alternates in sign on $x_1, \ldots, x_{n+1}$. Suppose there exists a polynomial $P(x;\bar{b})$ with

$$|f(x) - P(x;\bar{b})| < \rho$$

for every $x \in [A,B]$. Then between each pair of points x_k, x_{k+1} the two polynomials $P(x;\bar{a})$, $P(x;\bar{b})$ intersect. Hence, the polynomial

$$[f(x) - P(x;\bar{a})] - [f(x) - P(x;\bar{b})] = P(x;\bar{b}) - P(x;\bar{a})$$

has at least n distinct zeroes in $[A,B]$ which contradicts the assumption that $f_1, \ldots, f_n$ formed a T-system on $[A,B]$.

Next consider the necessity of the condition. Let $P(x;\bar{a})$ be the polynomial of least deviation and suppose it does not satisfy the given property. Then there exist k points $(k \leq n)$, $y_1 < y_2 < \cdots < y_k$ at which $|R(y_i)| = \rho_o$ and $R(y_i)$ alternates in sign. There then exist $k - 1$ points $x_1, x_2, \ldots, x_{k-1}$ at which $R(x_i) = 0$ with

$$A \leq y_1 < x_1 < y_2 < \cdots < x_{k-1} < y_k \leq B$$

and so that the inequalities

$$-\rho \leq f(x) - P(x;\bar{a}) < \rho - u$$

$$-\rho - u < f(x) - P(x;\bar{a}) \leq \rho$$

are satisfied in cyclic order in the intervals

$$[A,x_1],\ [x_1,x_2],\ \ldots\ [x_{k-1},B]$$

and for some u satisfying $0 < u < \rho/2$.

If $k = n$ then define

$$P(x;\bar{b}) = D(x,x_1,\ \ldots,\ x_{k-1})$$

and choose the sign of ε such that

$$\text{sign}\ [\varepsilon\ P(x;\bar{b})] = \text{sign}\ [f(y_1) - P(y_1;\bar{a})].$$

Then for $|\varepsilon|$ sufficiently small

$$|f(x) - P(x;\bar{a}) - \varepsilon\ P(x;\bar{b})| < \rho_o$$

for every $x \in [A,B]$. But this contradicts the assumption that $P(x;\bar{a})$ is a polynomial of least deviation.

If $k < n$, then let

$$m = n - k\ .$$

If m is even, choose a sufficiently small neighborhood, N, about x_{k-1} so that

$$|f(x) - P(x;\bar{a})| \leq \frac{\rho_o}{2}$$

for all $x \in N$. Then choose m points of N, $x_k,\ x_{k+1},\ \ldots,\ x_{n-1}$. If m is odd, choose $m - 1$ points $x_k,\ x_{k+1},\ \ldots,\ x_{n-2}$ in N and let $x_{n-1} = B$.

When m is even, there exists an ε for which

$$|f(x) - P(x;\bar{a}) - \varepsilon P(x;\bar{b})| < \rho_o$$

for every $x \in [A,B]$. When m is odd it is possible that at $x = B$

$$|f(B) - P(B;\bar{a}) - \varepsilon P(B;\bar{b})| = \rho_o$$

However, if another polynomial $P(x;\bar{c})$ is used where

$$P(B;\bar{c})\,[f(B) - P(B;\bar{a})] \geq 0,$$

then for both ε, ε_1 sufficiently small

$$|f(x) - P(x;\bar{a}) - \varepsilon P(x;\bar{b}) - \varepsilon_1 P(x;\bar{c})| < \rho_o$$

for $x \in [A,B]$. In either case a contradiction is obtained. This completes the proof.

CHAPTER III

BASIC ALGORITHMS

The first algorithm given is a variation of the second method of Remez as presented by E. L. Stiefel [9, 10] and called by him "the exchange method." It includes as a special case the method of E. N. Novodvarskii and I. Sh. Pinsker [12] as given by A. Schenitzer [13].

The outline of the theory, in particular the proof of the exchange theorem, given by Stiefel is constructive in the sense that it is a procedure for the solution of the uniform approximation problem. For this reason the theory will be developed first.

Let $f_1, \ldots, f_n$ form a T-system with respect to the closed and bounded set S and let $f \in C[S]$. Then the problem is to develop an algorithm which will determine a set of coefficients $a_1, \ldots, a_n$ which minimize

$$\max_{x \in S} \left| f(x) - \sum_{i=1}^{n} a_i f_i(x) \right| .$$

Let $X = \{x_1, \ldots, x_{n+1}\}$ be a set of $n + 1$ distinct points of S and let $P(x;\bar{a})$ be a polynomial on S. Then there exists a set of numbers $\{\alpha_1, \ldots, \alpha_{n+1}\}$ such that

$$\alpha_1 P(x_1;\bar{a}) + \alpha_2 P(x_2;\bar{a}) + \cdots + \alpha_{n+1} P(x_{n+1};\bar{a}) = 0 \qquad \text{(III-1)}$$

This follows from the fact that $f_1, \ldots, f_n$ satisfy Haar's condition on S and is easily verified by expanding the determinant

$$\begin{vmatrix} f_1(x_1) & \cdots & f_n(x_1) & P(x_1;\bar{a}) \\ \vdots & & \vdots & \vdots \\ f_1(x_{n+1}) & \cdots & f_1(x_{n+1}) & P(x_{n+1};\bar{a}) \end{vmatrix}$$

by the elements of the last column. Since the last column is a linear combination of the first n columns, the determinant is zero. Furthermore, each α_i is different from zero since the coefficients of Equation (III-1) as obtained by the expansion are the signed cofactors of the elements in the last column, which by Haar's condition must be nonzero.

Note also that the α_i are independent of the polynomial $P(x;\bar{a})$. Hence, (III-1) is satisfied for every $P(x;\bar{a})$.

Definition 11: A polynomial $P(x;\bar{a})$ is called a reference function with respect to the reference set X if the deviation $\rho_i = f(x_i) - P(x_i;\bar{a})$ corresponding to the elements $x_i \in X$ satisfy

$$\left.\begin{array}{l} \text{sign } \rho_i = \text{sign } \alpha_i \\ \text{sign } \rho_i = -\text{sign } \alpha_i \end{array}\right\} \quad \text{(III-2)}$$

or

for $i = 1,2, \ldots, n+1$.

The existence of reference functions is easily established. In fact, if ρ_i is the deviation of a polynomial $P(x;\bar{a})$ at the point $x_i \in X$, then a reference function can be obtained by solving the systems of equations on the following page for $b_1, \ldots, b_n$. Then $P(x;\bar{b})$ is a reference function with respect to X.

$$\begin{bmatrix} f_1(x_1) & \cdots & f_n(x_1) & |\rho_1| \text{ sign } \alpha_1 \\ \vdots & & \vdots & \vdots \\ f_1(x_{n+1}) & \cdots & f_n(x_{n+1}) & |\rho_{n+1}| \text{sign } \alpha_{n+1} \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_{n+1}) \end{bmatrix} \quad \text{(III-3)}$$

Definition 12: A reference function $P(x;\bar{b})$ is called a levelled reference function if the deviations ρ_i have the same absolute value at each point $x_i \in X$. The absolute value of ρ at these points is called the reference deviation.

A levelled reference function can be obtained from (III-3) by setting $|\rho_i| = \rho$ for $i = 1,2, \ldots, n+1$ and solving the resulting system for $b_1, \ldots, b_n, \rho$.

Suppose now that the equation

$$\rho_i = f(x_i) - P(x_i;\bar{a}) \quad \text{(III-4)}$$

is solved for $P(x_i;\bar{a})$ and substituted into (III-1). The result is

$$\sum_{i=1}^{n+1} \alpha_i [f(x_i) - \rho_i] = 0$$

or

$$\sum_{i=1}^{n+1} \alpha_i \rho_i = \sum_{i=1}^{n+1} \alpha_i f(x_i) . \quad \text{(III-5)}$$

Then if $P(x;\bar{a})$ is a reference function (III-5) becomes

$$\sum_{i=1}^{n+1} |\alpha_i| |\rho_i| = \pm \sum_{i=1}^{n+1} \alpha_i f(x_i) \quad \text{(III-6)}$$

From this we may conclude that the deviations of all reference functions corresponding to a fixed reference set (and function f) form a bounded set.

For a levelled reference function let

$$\rho_i = \rho \ \text{sign} \ \alpha_i$$

in (III-5) and solve for ρ. Then

$$\rho = \frac{\Sigma \alpha_i \ f(x_i)}{\Sigma |\alpha_i|} \tag{III-7}$$

Thus the value of the reference deviation for the levelled reference function can be obtained using (III-7) and polynomial interpolation. Note that the levelled reference function is not dependent on a particular reference function.

If (III-4) is solved for $f(x_i)$ and the result substituted into (III-7), one obtains

$$\rho = \frac{\Sigma \alpha_i [\rho_i + P(x_i;\bar{a})]}{\Sigma |\alpha_i|} ,$$

or

$$|\rho| = \frac{\Sigma |\alpha_i| \ |\rho_i|}{\Sigma |\alpha_i|} \tag{III-8}$$

But this implies that $|\rho|$ is a weighted average of the $|\rho_i|$ and thus

$$\min_i |\rho_i| \leq |\rho| \leq \max_i |\rho_i| \tag{III-9}$$

Therefore, if $P(x;\bar{a})$ is <u>any</u> reference polynomial with reference deviations ρ_i then from (III-8) it follows that

$$|\rho| \leq \max_i |\rho_i| . \qquad \text{(III-10)}$$

Let $P(x;\bar{a})$ be a levelled reference function with reference deviation ρ on the reference set X, and let $y \in S$ be such that

$$|f(y) - P(y;\bar{a})| = \max_{x \in S} |f(x) - P(x;\bar{a})| > \rho .$$

Then there exists an $x_i \in X$ such that if x_i is "exchanged" for y in X, $P(x;\bar{a})$ will be a reference function on the new reference set $X^* = X - \{x_i\} + \{y\}$. The proof of this will be given later in Theorem 9.

It follows from (III-8) that the new levelled reference functions $P(x;\bar{b})$ obtained by levelling $P(x;\bar{a})$ on the reference set X^* will have a reference deviation ρ^* which is strictly greater than ρ. Thus this process can be repeated and a sequence (which might terminate) of deviations $\{\rho_j^* : j = 1,2, \ldots\}$ will be obtained that satisfy the inequalities

$$\rho_1^* < \rho_2^* < \rho_3^* < \cdots .$$

Now suppose S is a discrete set of points $x_1, x_2, \ldots, x_m$ where $m > n+1$. Then by Theorem 5 there exists a polynomial $P(x;\bar{c})$ of least deviation from f on S. Furthermore, S being a discrete set implies that the sequence $\{\rho_j^*\}$ will terminate, since there are only a finite number of distinct reference sets X^*, and a reference

set cannot be repeated since $\rho_j^* < \rho_{j+1}^*$. Hence the last element λ of the sequence $\{\rho_j^*\}$ will satisfy

$$\lambda \leq \rho_M(f) \tag{III-11}$$

where

$$\rho_M(f) = \max_{x \in S} |f(x) - P(x;\bar{c})| .$$

If $\lambda < \rho_M(f)$ then there exists a levelled reference function $P(x;\bar{a})$ such that

$$\max_{x \in S} |f(x) - P(x;\bar{a})| = \lambda \tag{III-12}$$

since otherwise it would be possible to obtain a $\rho^* > \lambda$. But (III-12) would imply that $P(x;\bar{c})$ was not a polynomial of least deviation from f on S. Therefore, $\lambda = \rho_M(f)$ and the iterative procedure used to obtain λ converges to the polynomial of least deviation from f on S, subject of course to the proof of the exchange theorem.

This procedure applies equally well when S is a finite interval (see E. N. Novodvarskii and I. Sh. Pinsker [12]), but for the development of algorithms for numerical application the discrete case is sufficient. For S a countable compact set in E^r the proof has been given by E. W. Cheney and A. A. Goldstein [14].

The exchange theorem follows.

Theorem 9: Let $P(x;\bar{a})$ be a levelled reference function with reference deviation ρ on a reference set X. Suppose there exists a $y \in S$ such that

$$|f(y) - P(x;\bar{a})| > \rho .$$

Then there is an $x_i \in X$ such that, if y is exchanged for x_i, $P(x;\bar{a})$ will be a reference function on the nonreference set $X^* = X - \{x_i\} + \{y\}$.

Proof: Since $P(x;\bar{a})$ is a levelled reference function on $\{x_1, x_2, \ldots, x_n\}$ there exist numbers $\alpha_1, \ldots, \alpha_{n+1}$ which satisfy

$$\alpha_1 P(x_1;\bar{a}) + \cdots + \alpha_{n+1} P(x_{n+1};\bar{a}) = 0 \qquad \text{(III-13)}$$

and

$$\text{sign } \alpha_i = \pm \text{ sign } [f(x_i) - P(x_i;\bar{a})]$$

for $i = 1,2, \ldots, n+1$.

Next consider the set of points $\{x_1, \ldots, x_n, y\}$. $P(x;\bar{a})$ is a polynomial on S and hence there exists a set of numbers $\beta_1', \ldots, \beta_{n+1}'$ satisfying $\beta_i' \neq 0$ $(i = 1,2, \ldots, n+1)$ and

$$\beta_1' P(x_1;\bar{a}) + \cdots + \beta_n' P(x_n;\bar{a}) + \beta_{n+1}' P(y;\bar{a}) = 0,$$

for all polynomials. Then dividing by β_{n+1}' gives

$$\beta_1 P(x_1;\bar{a}) + \cdots + \beta_n P(x_n;\bar{a}) + 0 \cdot P(x_{n+1};\bar{a}) + P(y;a) = 0. \qquad \text{(III-14)}$$

$$(\beta_1 - \theta\alpha_1) P(x_1,\bar{a}) + \cdots + (\beta_n - \theta\alpha_n) P(x_n;\bar{a}) - \theta \alpha_{n+1} P(x_{n+1},\bar{a}) + P(y;\bar{a}) = 0$$

by subtracting from III-14, θ times III-13.

If $f(y) - P(y;\bar{a}) > 0$ and $\operatorname{sign} \alpha_i = \operatorname{sign} \{f(x_i) - P(x_i;\bar{a})\}$

or if $f(y) - P(y;a) < 0$ and $\operatorname{sign} \alpha_i = -\operatorname{sign} \{f(x_i) - P(x_i;\bar{a})\}$

then define

$$\theta_o = \min \begin{cases} \min \dfrac{\beta_i}{\alpha_i}, & i = 1,2, \ldots, n \\ 0 \end{cases} \qquad \text{(III-15)}$$

If $\theta_o = 0$ then $\operatorname{sign} \beta_i = \operatorname{sign} \alpha_i$ and replace x_{n+1} by y. If $\theta_o = \beta_{i_o}/\alpha_{i_o}$ then

$$\frac{\beta_{i_o}}{\alpha_{i_o}} < \frac{\beta_i}{\alpha_i} \qquad \text{for } i = 1,2, \ldots, n, \; i_o \neq i$$

and $\operatorname{sign} (\beta_i - \theta_o \alpha_i) = \operatorname{sign} \alpha_i$, $i \neq i_o$ with $\beta_{i_o} - \theta_o \alpha_{i_o} = 0$. Thus replace x_{i_o} by y.

On the other hand if $f(y) - P(y;\bar{a}) > 0$ and $\operatorname{sign} \alpha_i = -\operatorname{sign} \{f(x_i) - P(x_i;\bar{a})\}$ or if $f(y) - P(y;\bar{a}) < 0$ and $\operatorname{sign} \alpha_i = \operatorname{sign} \{f(x_i) - P(x_i;\bar{a})\}$ then define

$$\theta_o = \max \begin{cases} \max \dfrac{\beta_i}{\alpha_i}, & i = 1,2, \ldots, n \\ 0 \end{cases} \qquad \text{(III-16)}$$

If $\theta_o = 0$ then $\operatorname{sign} \beta_i = -\operatorname{sign} \alpha_i$ and replace x_{n+1} by y. If $\theta_o = \beta_{i_o}/\alpha_{i_o}$ then

$$\frac{\beta_{i_o}}{\alpha_{i_o}} > \frac{\beta_i}{\alpha_i} \qquad \text{for } i = 1,2, \ldots, n, \; i \neq i_o$$

and sign $\alpha_i = -$ sign $\{\beta_i - \theta_o \alpha_i\}$ with $\beta_{i_o} - \theta_o \alpha_{i_o} = 0$. Thus in this case replace x_{i_o} by y. This completes the proof.

It was assumed earlier in the proof of the convergence of the sequence $\{\rho_k\}$ that S was a discrete set of points or had been replaced by a discrete set. In what follows it will be assumed that S is a discrete set of points unless stated otherwise.

In order to give an algorithm based on the exchange theorem, a starting procedure is needed to obtain a levelled reference function and the numbers $\alpha_1, \ldots, \alpha_{n+1}$. The first step is to select a set of $n + 1$ points of S. The best choice of these $n + 1$ points would be a set which yielded the largest reference deviation ρ for the starting reference function. In general a method of determining such a reference is not known. However, a good <u>a priori</u> choice when S is an interval is given by the local extreme points of the Tchebysheff polynomial $T_n(x)$ on S. For the internal $[-1,1]$ these are defined by

$$x_i = -\cos \frac{i\pi}{n}, \qquad i = 0,1, \ldots, n . \tag{iii-17}$$

A heuristic motivation for such a choice can be made when $f_1(x) = 1$, $f_2(x) = x$, ..., $f_n(x) = x^{n-1}$. Suppose $f(x)$ is a polynomial of degree n (or at least can be closely approximated by such a polynomial). Then the error function at least deviation

$$R(x) = f(x) - P_n(x)$$

is proportional to the Tchebysheff polynomial of degree n which assumes its maximum value at these $n + 1$ points. However, it must be pointed out that for S a discrete set this choice may not be possible. In this case choose those points nearest the Tchebysheff abscissae.

This choice was used by W. Fraser and J. F. Hart [15] in their work with rational approximations using an algorithm similar to Stiefel's exchange theorem. They reported that good results were obtained even though the above argument does not apply for rational functions.

With the initial x_i's chosen the α_i's can be obtained by setting $\alpha_{n+1} = 1$ and solving the system of equations,

$$\begin{bmatrix} f_1(x_1) & \cdots & f_1(x_n) \\ \vdots & & \vdots \\ f_n(x_1) & \cdots & f_n(x_n) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} -f_1(x_{n+1}) \\ \vdots \\ -f_n(x_{n+1}) \end{bmatrix} \qquad \text{(III-18)}$$

for $\alpha_1, \ldots, \alpha_n$.

Equation (III-18) has a unique nontrivial solution since $f_1, \ldots, f_n$ is a T-system on $\{x_1, \ldots, x_m\}$. Furthermore each $\alpha_i \neq 0$; hence

$$\sum_{i=1}^{n+1} |\alpha_i| \neq 0 .$$

Next let $\lambda_i = \operatorname{sign} \alpha_i$ $(i = 1,2, \ldots, n+1)$ and consider the system of equations on the following page.

$$\begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 f_1(x_1) & \cdots & \lambda_{n+1} f_1(x_{n+1}) \\ \vdots & & \vdots \\ \lambda_1 f_n(x_1) & \cdots & \lambda_{n+1} f_n(x_{n+1}) \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{n+1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \text{(III-19)}$$

Letting

$$\gamma_j = \frac{|\alpha_j|}{\sum_{i=1}^{n+1} |\alpha_i|}, \qquad (j = 1,2, \ldots, n+1) \qquad \text{(III-20)}$$

it follows that

$$\Sigma \gamma_j = \sum_{j=1}^{n+1} \left\{ \frac{|\alpha_j|}{\sum_{i=1}^{n+1} |\alpha_i|} \right\} = 1 .$$

Also

$$\sum_{j=1}^{n+1} \gamma_j \lambda_j f_i(x_j) = \frac{1}{\sum_{k=1}^{n+1} |\alpha_k|} \left\{ \sum_{j=1}^{n+1} \lambda_j |\alpha_j| f_i(x_j) \right\} = 0 .$$

Therefore

$$\gamma_j = \frac{|\alpha_j|}{\sum_{i} |\alpha_i|} , \qquad (j = 1,2, \ldots, n)$$

is a solution to (III-19).

It can now be shown that the $\gamma_1, \ldots, \gamma_{n+1}$ are unique and therefore the matrix in (III-19) is nonsingular. To do this assume $\beta_1, \ldots, \beta_{n+1}$ is another solution to (III-19) such that at least one β_i, say, β_{i_o} satisfies

$$\beta_{i_o} \neq \gamma_{i_o} .$$

Substituting β_i, γ_i into (III-19) gives

$$\sum \beta_i \, [\lambda_i \, P(x_i;\bar{a})] = 0 \qquad \text{(III-21)}$$

and

$$\sum \gamma_i \, [\lambda_i \, P(x_i;\bar{a})] = 0 \qquad \text{(III-22)}$$

for any polynomial $P(x;\bar{a})$.

Then multiplying (III-21) by γ_{i_o} and (III-22) by β_{i_o} and subtracting one obtain

$$(\gamma_1\beta_{i_o} - \gamma_{i_o}\beta_1)\,(\lambda_1 P(x_1;\bar{a})) + \cdots$$

$$+ (\gamma_{n+1}\beta_{i_o} - \gamma_{i_o}\beta_{n+1})\,(\lambda_{n+1}P(x_{n+1};\bar{a})) = 0 .$$

Note that the coefficient of $\lambda_{i_o} P(x_{i_o};\bar{a})$ is zero in this equation. Therefore all the coefficients must be zero since otherwise $f_1, \ldots, f_n$ would not be a T-system on $\{x_1, \ldots, x_m\}$. Thus

$$\beta_i = \frac{\beta_{i_o}}{\gamma_{i_o}} \gamma_i , \qquad (i = 1,2, \ldots, n+1)$$

and hence

$$\sum_{i=1}^{n+1} \beta_i = \frac{\beta_{i_o}}{\gamma_{i_o}} \sum_{i=1}^{n+1} \gamma_i = \frac{\beta_{i_o}}{\gamma_{i_o}} .$$

But $\sum_{i=1}^{n+1} \beta_i = 1$ by Equation (III-19) from which it follows that

$$\beta_i = \gamma_i , \qquad (i = 1,2, \ldots, n+1) .$$

From this it is easily seen that the matrix

$$B = \begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 f_1(x_1) & \cdots & \lambda_{n+1} f_1(x_{n+1}) \\ \vdots & & \vdots \\ \lambda_1 f_n(x_1) & \cdots & \lambda_{n+1} f_n(x_{n+1}) \end{bmatrix}$$

is nonsingular.

It would be possible now to present an algorithm based on the Exchange Theorem. However, for notational purposes the Exchange Theorem will be reformulated before giving the algorithm.

Let $\sigma = \text{sign}\{f(y) - P(y,\bar{a})\}$ and let $\gamma_1, \gamma_2, \cdots, \gamma_{n+1}$ be the solution of the linear system

$$\begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 f_1(x_1) & \cdots & \lambda_{n+1} f_1(x_{n+1}) \\ \vdots & & \vdots \\ \lambda_1 f_n(x_1) & \cdots & \lambda_{n+1} f_n(x_{n+1}) \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{n+1} \end{bmatrix} = \begin{bmatrix} 1 \\ \sigma f_1(y) \\ \vdots \\ \sigma f_n(y) \end{bmatrix}$$

where $\lambda_i = \text{sign}\ \alpha_i$. Thus

$$\gamma_1\lambda_1 P(x_1;\bar{a}) + \cdots + \gamma_{n+1}\lambda_{n+1} P(x_{n+1};\bar{a}) = \sigma\, P(y;\bar{a}) \qquad \text{(III-23)}$$

holds for all polynomials, in particular for $P(x;\bar{a})$. Equation (III-13) can be written in the form

$$|\alpha_1|\ \lambda_1 P(x_1;\bar{a}) + \cdots + |\alpha_{n+1}|\ \lambda_{n+1} P(x_{n+1};\bar{a}) = 0\,. \qquad \text{(III-24)}$$

Now choose i_o so that

$$\frac{\gamma_{i_o}}{|\alpha_{i_o}|} = \max \left\{ \frac{\gamma_i}{|\alpha_i|} : i = 1,2, \ldots, n+1 \right\} .$$

Multiplying (III-24) by γ_{i_o} and (III-23) by α_{i_o} and then substituting the resulting equations gives

$$\begin{aligned}
&(\gamma_{i_o}|\alpha_1) - |\alpha_{i_o}|\ \gamma_1)\ \lambda_1 P(x_1,\bar{a}) + \cdots + (\gamma_{i_o}|\alpha_{i_o-1}| \\
&\qquad - |\alpha_{i_o}|\ \gamma_{i_o-1})\ \lambda_{i_o} P(x_{i_o-1};\bar{a}) + \sigma\ \lambda_{i_o} P(x_{i_o};\bar{a}) \\
&\qquad + (\gamma_{i_o}|\alpha_{i_o+1}| - |\alpha_{i_o}|\ \gamma_{i_o+1})P(x_{i_o+1};\bar{a}) + \cdots \\
&\qquad + (\gamma_{i_o}|\alpha_{n+1}| - |\alpha_{i_o}|\gamma_{n+1})\ \lambda_{n+1}P(x_{n+1};\bar{a}) \\
&\qquad + \sigma|\alpha_{i_o}|P(y,\bar{a}) = 0 . \qquad \text{(III-25)}
\end{aligned}$$

From

$$0 \leq \frac{\gamma_{i_o}}{|\alpha_{i_o}|} - \frac{\gamma_i}{|\alpha_i|} = \frac{\gamma_{i_o}|\alpha_i| - \gamma_i|\alpha_{i_o}|}{|\alpha_{i_o}| \cdot |\gamma_i|}$$

it follows that

$$0 \leq \gamma_{i_o}|\alpha_i| - \gamma_i|\alpha_{i_o}| \quad \text{for} \quad i = 1,2, \ldots, n+1 .$$

Since III-25 holds not only for $P(x;\bar{a})$, but for all polynomials we see that

$$0 < \gamma_{i_o}|\alpha_i| - \gamma_i|\alpha_{i_o}| \quad \text{for} \quad i \neq i_o .$$

It follows that $P(x;\bar{a})$ is a reference polynomial on the reference set $X^* = X - \{x_{i_o}\} + \{y\}$, by observing that the coefficients of $P(x_i;\bar{a})$ in (III-25) have the proper signs.

An algorithm based on the Exchange Theorem follows.

Algorithm 1:

Input: $x_1, x_2, \ldots, x_m, \quad m > n$

$f(x_1), f(x_2), \ldots, f(x_m)$

$I = \{i_1, i_2, \ldots, i_{n+1}\}$ which we shall for simplicity denote by $\{1,2, \ldots, n+1\}$

$\alpha_i \quad$ for $\quad k = 1,2, \ldots, n+1$

Step 1.. Set $\lambda_k = \text{sign } \alpha_i \quad (k = 1,2, \ldots, n+1)$ and compute $\rho_o, a_1, \ldots, a_n$ using

$$(\rho_o, a_1, \ldots, a_n) \begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 f_1(x_1) & \cdots & \lambda_{n+1} f_1(x_{n+1}) \\ \vdots & & \vdots \\ \lambda_1 f_m(x_1) & \cdots & \lambda_{n+1} f_n(x_{n+1}) \end{bmatrix} = (\lambda_1 f(x_1), \ldots, \lambda_{n+1} f(x_{n+1})).$$

Step 2.. Determine j_o such that

$$|R(x_{j_o})| = \max_j |R(x_j)|, \quad (j = 1,2, \ldots, m).$$

If

$$|R(x_{j_o})| = |\rho_o|$$

then the polynomial obtained in Step 1 is the polynomial of least deviation from f on $\{x_1, \ldots, x_m\}$ and the algorithm stops.

Otherwise proceed to Step 3.

Step 3.. Set $\sigma = \text{sign } R(x_{j_o})$ and compute $\gamma_1, \ldots, \gamma_{n+1}$ using

$$\begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 f_1(x_1) & \cdots & \lambda_{n+1} f_1(x_{n+1}) \\ \vdots & & \vdots \\ \lambda_1 f_n(x_1) & \cdots & \lambda_{n+1} f_n(x_{n+1}) \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{n+1} \end{bmatrix} = \begin{bmatrix} 1 \\ \sigma f_1(x_{j_o}) \\ \vdots \\ \sigma f_n(x_{j_o}) \end{bmatrix},$$

Step 4.. Select i_o so that

$$\theta = \frac{\gamma_{i_o}}{|\alpha_{i_o}|} = \max \left\{ \frac{\gamma_i}{|\alpha_i|} : (i = 1,2, \ldots, n+1) \right\}$$

Step 5.. Replace i_o by j_o in I, and compute

$$\alpha_i = \lambda_i(\theta|\alpha_i| - \gamma_i), \qquad i \neq i_o$$

and

$$\alpha_{j_o} = \sigma\theta .$$

Then return to Step 1.

A close examination of Algorithm 1 with the simplex procedure of linear programming in mind leads one to believe that it might be simply a special case of the linear programming formulation of the uniform approximation problem. It was for this reason that the Exchange

Theorem was reformulated before giving the algorithm. In order to pursue this idea further the development of the simplex algorithm will be given next. It is assumed here that the reader is familiar with the theory of linear programming as it is presented in S. Gass [16].

Let $f_1, \ldots, f_n$ form a T-system with respect to a finite point set $S = \{x_1, x_2, \ldots, x_m\}$, and let $f \in C(S)$. Then the uniform approximation problem is to determine numbers $a_1, \ldots, a_n$ such that

$$\rho(f) = \max_j |f(x_j) - P(x_j;\bar{a})| \qquad \text{(III-26)}$$

is minimized.

Now, (III-26) can be replaced by the inequalities

$$-\rho \leq f(x_j) - P(x_j;\bar{a}) \leq \rho, \quad (j = 1,2, \ldots, m)$$

or

$$\begin{aligned} \rho + P(x_j;\bar{a}) &\geq f(x_j) \\ \rho - P(x_j;\bar{a}) &\geq -f(x_j) \end{aligned} \quad , \quad (j = 1,2, \ldots, m) \qquad \text{(III-27)}$$

With the use of (III-27) the problem can be expressed in the standard linear programming form.

Form 1: Minimize ρ subject to the constraints

$$\begin{bmatrix} 1 & f_1(x_1) & \cdots & f_n(x_1) \\ 1 & -f_1(x_1) & \cdots & -f_n(x_1) \\ \vdots & \vdots & & \vdots \\ 1 & f_1(x_m) & \cdots & f_n(x_m) \\ 1 & -f_1(x_m) & \cdots & -f_n(x_m) \end{bmatrix} \begin{bmatrix} \rho \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \geq \begin{bmatrix} f(x_1) \\ -f(x_1) \\ \vdots \\ f(x_m) \\ -f(x_m) \end{bmatrix} \qquad \text{(III-28)}$$

The dual of this program is:

Form II: Maximize $\Sigma \alpha_i c_i$ subject to the constraints

$$\alpha_i \geq 0, \quad (i = 1,2, \ldots, 2m) \tag{III-29}$$

and

$$\begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ f_1(x_1) & -f_1(x_1) & \cdots & f_1(x_m) & -f_1(x_m) \\ \vdots & \vdots & & \vdots & \vdots \\ f_n(x_1) & -f_n(x_1) & \cdots & f_n(x_m) & -f_n(x_m) \end{bmatrix} \begin{bmatrix} \alpha \\ \vdots \\ \alpha_{2m} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{III-30}$$

where

$$\bar{c} = [f(x_1),-f(x_1), \ldots, f(x_m),-f(x_m)] .$$

The standard simplex algorithm will be applied to Form II of the linear program.

Definition 13: A feasible solution to the linear programming problem is a vector $\bar{\alpha}$ which satisfies the constraints; i.e., $\bar{\alpha}$ is a feasible solution if it satisfies (III-29) and (III-30).

Definition 14: A basic feasible solution is a feasible solution with no more than $n + 1$ nonzero elements.

In order to use the simplex procedure a basic feasible solution must be known; i.e., a starting procedure is needed just as in Algorithm 1 to obtain a basic feasible solution. Such procedures are well known and may be found in S. Gass [16]. The most general of these is the one using an "artificial basis."

Since the principal concern here is the algorithm, it is assumed that a basic feasible solution is given.

Denote Equation (III-30) by

$$A\,\bar{\alpha} = \bar{b}$$

and let A_j be the j^{th} column of A. Let the nonzero elements of the given basic feasible solution be given by the vector

$$\bar{\beta}_o = (\beta_1, \beta_2, \ldots, \beta_{n+1})^T$$

where β_k corresponds to the i_k^{th} column of A. Let

$$B = \begin{bmatrix} | & | & & | \\ A_{i_1} & A_{i_2} & \cdots & A_{i_{n+1}} \\ | & | & & | \end{bmatrix}.$$

Thus

$$\bar{\beta}_o = B^{-1}\,\bar{b}\,.$$

The simplex algorithm follows. The objective function will be changed to: Minimize $-\Sigma\,\alpha_i\,c_i$.

Algorithm 2:

Input: $\bar{c} = [-f(x_1), f(x_1), \ldots, -f(x_m), f(x_m)]$

$A_j \quad j = 1,2, \ldots, 2m$

$I = \{i_1, i_2, \ldots, i_{n+1}\}, \ \bar{\beta}_o$

Step 1.. For each $j \notin I$ compute

$$B^{-1} A_j = \bar{\beta}_j = \begin{bmatrix} \beta_{j_1} \\ \vdots \\ \beta_{j_{n+1}} \end{bmatrix}.$$

Step 2.. Compute $z_j - c_j$ for $j \notin I$ where

$$z_j = \sum_{k=1}^{n+1} \beta_{jk} c_{i_k}$$

and c_{i_k} is the i_k^{th} component of the vector $\bar{c}$.

Step 3.. If $z_j - c_j \leq 0$ for $j = 1,2, \ldots, 2m$ then $\bar{\beta}_o$ is the minimum feasible solution. Otherwise proceed to Step 4.

Step 4.. Determine a j_o such that

$$z_{j_o} - c_{j_o} = \max_j (z_j - c_j) .$$

Step 5.. Determine an i_o such that

$$\frac{\beta_{i_o}}{\beta_{j_o i_o}} = \min \left\{ \frac{\beta_i}{\beta_{j_o i}} : \beta_{j_o i} > 0 \right\}$$

where β_i is the i^{th} component of $\bar{\beta}_o$.

Step 6.. Replace i_o by j_o in I and A_{i_o} by A_{j_o}, and return to Step 1.

The comparison of Algorithms 1 and 2 is now possible. Assume that a set of $n + 1$ points $\{x_1, \ldots, x_{n+1}\}$ is given and that the α_i are computed using (III-18). If the indices $I = \{i_1, i_2, \ldots, i_{n+1}\}$ are chosen so that

$$A_{i_j} = \{1, \lambda_j f_1(x_j), \ldots, \lambda_j f_n(x_j)\}^T, \quad (j = 1,2, \ldots, n+1) ;$$

then the matrix

$$B = \begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 f_1(x_1) & \cdots & \lambda_{n+1} f_1(x_{n+1}) \\ \vdots & & \vdots \\ \lambda_1 f_n(x_1) & \cdots & \lambda_{n+1} f_n(x_{n+1}) \end{bmatrix}$$

is the same as the matrix used in Step 1 of Algorithm 1. Furthermore, B is nonsingular and

$$\bar{\beta}_o = B^{-1} \bar{b}$$

is a basic feasible solution of the linear program. This follows from the development of the solution of Equation (III-19). Hence both procedures determine the same polynomial $P(x;\bar{a})$.

Next consider the selection of j_o. In the linear programming algorithm j_o is determined by

$$z_{j_o} - c_{j_o} = \max_j (z_j - c_j) ,$$

where

$$z_j - c_j = \sum_{k=1}^{n+1} \beta_{j_k} c_k - c_j ,$$

$$c_k = -\lambda_k f(x_k), \quad (k = 1,2, \ldots, n+1),$$

and

$$\sum_{k=1}^{n+1} \beta_{j_k} \lambda_k P(x_k;\bar{a}) = \sigma P(x_j;\bar{a}).$$

In this last equation $\sigma = \pm 1$ depending on the selection of A_j, i.e.,

$$A_j = (1, \sigma f_1(x_j), \ldots, \sigma f(x_j))^T.$$

Thus

$$z_j - c_j = -\sum_k \beta_{j_k} \lambda_k f(x_k) - \sigma(-f(x_j))$$

or

$$z_j - c_j = -\sum_k \beta_{j_k} [\lambda_k f(x_k)] + \sigma f(x_j).$$

But

$$\lambda_k f(x_k) - \lambda_k P(x_k;\bar{a}) = \rho > 0$$

and

$$f(x_j) - P(x_j;\bar{a}) = \sigma\rho_j > 0,$$

where

$$\rho_j = |f(x_j) - P(x_j;\bar{a})|.$$

Hence

$$z_j - c_j = -\sum \beta_{j_k} [\rho + \lambda_k P(x_k;\bar{a})] + \sigma[\sigma\rho_j + P(x_j;\bar{a})]$$

$$= -\rho \sum \beta_{j_k} - \sum \beta_{j_k} \lambda_k P(x_k;\bar{a}) + \rho_j + \sigma P(x_j;\bar{a}).$$

Now, $\sum_{k=1}^{n+1} \beta_{j_k} = 1$ and $\sum_{k=1}^{n+1} \beta_{j_k} \lambda_k P(x_k;\bar{a}) = \sigma P(x_j;\bar{a})$ and

therefore

$$z_j - c_j = -\rho - \sigma P(x_j;\bar{a}) + \rho_j + \sigma P(x_j;\bar{a})$$

or

$$z_j - c_j = \rho_j - \rho .$$

Hence,

$$\max_j (z_j - c_j) = \max_j (\rho_j - \rho)$$

where

$$\rho_j = |f(x_j) - P(x_j;\bar{a})| .$$

Thus j_o is the same in both algorithms.

The comparison will be complete if it can be shown that the determination of i_o is the same in both algorithems. To see this, multiply Equation (III-25) by $1/\gamma_{i_o}$ to obtain

$$(|\alpha_1| - \theta^*\gamma_1)\lambda_1 P(x_1;\bar{a}) + \cdots + (|\alpha_{i_o-1})| - \theta^*\gamma_{i_o-1})\lambda_{i_o-1} P(x_{i_o-1};\bar{a})$$

$$+ (|\alpha_{i_o+1}| - \theta^*\gamma_{i_o+1})\lambda_{i_o+1} P(x_{i_o+1};\bar{a}) + \cdots + (|\alpha_{n+1}|$$

$$- \theta^*\gamma_{n+1})\lambda_{n+1} P(x_{n+1};\bar{a}) + \sigma\theta^* P(y;\bar{a}) = 0 ,$$

where

$$\theta^* = \frac{|\alpha_{i_o}|}{\gamma_{i_o}} .$$

It should be noted that $\gamma_{i_o} \neq 0$ because of the requirements

$$\sum \gamma_i = 1 \quad \text{and} \quad \frac{\gamma_{i_o}}{|\alpha_{i_o}|} = \max_j \frac{\gamma_i}{|\alpha_i|} .$$

Next, comparing the equations of Step 1 of the two algorithms and observing the starting procedures, it is seen that

$$|\alpha_i| = \beta_i$$

and

$$\gamma_i = \beta_{j_o i} .$$

Thus

$$\max \frac{\gamma_i}{|\alpha_i|} = \min_j \left\{ \frac{\beta_i}{\beta_{j_o i}} : \beta_{j_o i} > 0 \right\}$$

$$= \min_j \left\{ \frac{|\alpha_i|}{\gamma_i} : \gamma_i > 0 \right\}$$

and hence i_o is the same in both algorithms.

A close examination of the Exchange Theorem will show that the choice of j_o by

$$|R(x_{j_o})| = \max_j |R(x_j)|$$

does not necessarily yield the maximum increase in the reference deviation ρ. But in view of the equivalence of the two algorithms this is to be expected since

$$\max_j (z_j - c_j)$$

does not yield the maximum increase in the objective function for the simplex algorithm. If the two algorithms were modified to obtain this maximum increase the resulting algorithms would still be equivalent, but

the number of computations necessary during each iteration would be sharply increased. Experience with the simplex procedure indicates that this modification does not, in general, decrease the number of iterations sufficiently to be of much practical value, Gass [16].

CHAPTER IV

MODIFICATION OF THE BASIC ALGORITHMS

The first modification considered concerns the starting procedure for Algorithm 1. It is due to E. W. Cheney and A. A. Goldstein [14]. The procedure which they use chooses a set of points $\{x_1, \ldots, x_n\}$ and adds an "artificial vector" constructed from these n points to complete the basis. It does not require the initial computation of the $\alpha_1, \ldots, \alpha_{n+1}$. The matrix used is

$$\begin{bmatrix} 1 & \cdots & 1 & 1 \\ f_1(x_1) & \cdots & f_1(x_n) & -\sum_{j=1}^{n} f_1(x_j) \\ \vdots & & \vdots & \vdots \\ f_n(x_1) & \cdots & f_n(x_n) & -\sum_{j=1}^{n} f_n(x_j) \end{bmatrix}, \qquad \text{(IV-1)}$$

the artificial vector being

$$A_{n+1} = \left[1, -\sum_{j=1}^{n} f(x_j), \ldots, -\sum_{j=1}^{n} f_n(x_j)\right]^T. \qquad \text{(IV-2)}$$

Associated with this artificial vector is a number b which takes the place of $f(x_{n+1})$, so that if the algorithm converges with the artificial vector still in the basis then b is increased and the algorithm is started over. This process is repeated until the algorithm converges with the artificial vector removed. The paper shows that if b is

increased indefinitely then the $\max_{x \in S} |R(x)|$ is unbounded and no solution exists.

While this starting procedure has the advantage that the programming necessary to compute $\alpha_1, \ldots, \alpha_{n+1}$ is not needed, additional computation would be needed to remove the artificial vector. Furthermore, if it becomes necessary to increase b, then the additional computation would exceed the computation to obtain $\alpha_1, \ldots, \alpha_{n+1}$, since then the algorithm must be started over from the first step.

The second modification is obtained by restricting the set S to be a finite interval, $[c,d]$. Then the set of points $\{x_1, \ldots, x_m\}$ may be indexed so that

$$c \leq x_1 < x_2 < \cdots < x \leq d .$$

Thus applying Theorem 8 it follows that the polynomial of least deviation from f on S is characterized by the property that the error function $R(x)$ alternates in sign on $n + 1$ points of S and assumes its maximum on these $n + 1$ points. Furthermore, if $\{x_1, \ldots, x_{n+1}\}$ is any ordered set of $n+1$ points of S, then the polynomial of least deviation from f on $\{x_1, \ldots, x_{n+1}\}$ is uniquely characterized by the same properties (see Theorem 7).

Thus for every ordered set of $n + 1$ points $\{x_1, \ldots, x_{n+1}\}$, the polynomial of least deviation from f on these $n + 1$ points satisfies

$$\text{sign } R(x_i) = - \text{ sign } R(x_{i+1}) \qquad \text{(IV-3)}$$

and

$$|R(x_i)| = \text{constant}, \quad (i = 1,2, \ldots, n+1) .$$

But

$$\text{sign } \alpha_i \equiv \lambda_i = \pm \text{ sign } R(x_i), \quad (i = 1,2, \ldots, n+1)$$

and hence Equation (IV-3) implies

$$\text{sign } \alpha_i = - \text{ sign } \alpha_{i+1}.$$

Therefore the matrix given in Step 1 of Algorithm 1 becomes

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ f_1(x_{i_1}) & - f_1(x_{i_{1+1}}) & \cdots & (-1)^n f_1(x_{i_{n+1}}) \\ \vdots & \vdots & & \vdots \\ f (x_{i_1}) & - f_n(x_{i_{1+1}}) & \cdots & (-1)^n f_1(x_{i_{n+1}}) \end{bmatrix} .$$

Hence it is no longer mandatory to have a starting procedure to compute $\alpha_1, \ldots, \alpha_{n+1}$.

In addition, the determination of the i_o to be replaced by j_o as in Step 4 can be changed to the selection of an i_o such that when j_o is replaced by i_o the error function $R(x)$ alternates in sign on the resulting set of points.

The algorithm which results from these changes is a modification of the one given by A. Shenitzer [13].

Algorithm 3:

Input: x_i, $f(x_i)$, $(i = 1,2, \ldots, m)$

$I = (1,2, \ldots, n+1)$

Step 1.. Solve

$$(\rho, a_1, \ldots, a_n) \begin{bmatrix} 1 & -1 & \cdots & (-1)^n \\ f_1(x_1) & f_1(x_2) & \cdots & f_1(x_{n+1}) \\ \vdots & \vdots & & \vdots \\ f_n(x_1) & f_n(x_2) & \cdots & f_n(x_{n+1}) \end{bmatrix} = [f(x_1) \cdots f(x_{n+1})]$$

for ρ, a_1, $\ldots$, a_n.

Step 2.. Determine j_o so that

$$|R(x_{j_o})| = \max_j |R(x_j)| .$$

Step 3.. If $|R(x_{j_o})| = \max \{|R(x_j)| : j \in I\}$, then $P(x;\bar{a})$ is the polynomial of least deviation from f on $\{x_1, \ldots, x_m\}$. Otherwise proceed to Step 4.

Step 4.. For convenience in notation assume $I = \{1,2, \ldots, n+1\}$. Then, if there exists an i, $i + 1 \in I$ such that $x_{j_o} \in (x_i, x_{i+1})$, replace i by j_o in I when $\operatorname{sign} R(x_{j_o}) = \operatorname{sign} R(x_i)$; otherwise replace $i + 1$ by j_o. If $x_{j_o} \in [c, x_1]$, then replace 1 by j_o in I when $\operatorname{sign} R(x_{j_o}) = \operatorname{sign} R(x_1)$; otherwise replace $n + 1$ by j_o. If $x_{j_o} \in (x_{n+1}, d]$, then replace $n + 1$ by j_o in I when

sign $R(x_{j_o}) = \text{sign } R(x_{n+1})$; otherwise replace 1 by j_o.

An alternate approach to the uniform approximation problem is to solve the system of equations of Step 1, Algorithm 3 for ρ in terms of $x_1, \ldots, x_{n+1}$ using Cramer's rule. The problem then becomes one of maximizing a function of $n + 1$ variables. This approach for the case in which $f_i(x) = x^{i-1}$, $i = 1,2, \ldots, n+1$ has been developed and applied by P. C. Curtiss, Jr., and W. L. Frank [17]. A restricted version of their method is given below in Algorithm 4. An outline of the theory follows.

Let f be a continuous function defined on a finite interval $[c,d]$. Let $f_i = x^{i-1}$ so that f is to be approximated by the functions $1,x,x^2, \ldots, x^n$. Then replace $[c,d]$ by an ordered set of points $\{x_1, x_2, \ldots, x_m\}$; i.e., $x_1 < x_2 < \cdots < x_m$. Assuming $i = 0,1, \ldots,n+1$ the system of equations for $\rho, a_0,\ldots,a_n$ becomes

$$a_o + a_1 x_j + a_2 x_j^2 + \cdots + a_n x_j^n + (-1)^j \rho = f(x_j), \quad \text{(IV-4)}$$

$$j = 1,2, \ldots, n+2 .$$

Now, suppose (IV-4) is solved for ρ using Cramer's rule. Then

$$\rho(x_o, \ldots, x_{n+1}) = \frac{\sum_{j=o}^{n+1} (-1)^j f(x_j) C_j(x_o, \ldots, x_{n+1})}{\sum_{j=o}^{n+1} C_j(x_o, \ldots, x_{n+1})}$$

where

$$C_j(x_o, \ldots, x_{n+1}) = \sum_{\substack{i=o \\ i \neq j}}^{n+1} |x_i - x_j|^{-1} .$$

Then, if a subset $y_o, y_1, \ldots, y_{n+1}$ of $\{x_1, \ldots, x_m\}$ is determined so that $\rho(y_o, \ldots, y_{n+1})$ is maximum over all possible subsets of $n + 2$ elements of $\{x_1, \ldots, x_m\}$, solving the system of equations in (IV-4) using $y_o, \ldots, y_{n+1}$ for the x_j^2 will yield the polynomial of least deviation from f on $\{x_1, \ldots, x_m\}$. The proof of this last result is given in the paper by P. C. Curtiss, Jr., and W. L. Frank [17].

In the exchange process of Algorithm 3 it was necessary to treat as a special case the exchange of j_o for some $i \leq I$ when x_{j_o} was contained in a one-sided neighborhood of an endpoint (i.e., when $x_{j_o} \in [c, x_1]$ or $(x_{n+1}, d]$). A similar treatment would be necessary here. However, if it is assumed that $f(x)$ is differentiable and that its $(n + 1)^{st}$ derivative does not change sign, the endpoints of the interval must be contained in the set $\{y_o, \ldots, y_{n+1}\}$ which maximizes ρ (see Hart and Fraser [15]). Thus the special treatment at the endpoints is not necessary. It must be emphasized that this assumption is not necessary and that it is being made in order to simplify the statement of the algorithm.

Let $y_1, \ldots, y_{n+2}$ be $n + 2$ distinct elements of $\{x_1, \ldots, x_m\}$ and define

$$\rho_k(x) = \rho(y_1, \ldots, y_{k-1}, x, y_{k+1}, \ldots, y_{n+2})$$

where

$$y_{k-1} < x < y_{k+1}$$

for $k = 2,3, \ldots, n+1$. With this definition the algorithm can now be given.

Algorithm 4:

Input: $x_i, f(x_i), \quad (i = 1,2, \ldots, m)$

$I = (1,2, \ldots, n+2)$ so that $y_i = x_i$, $(i = 1,2, \ldots, n+2)$.

Step 1.. Set $k = 2$.

Step 2.. Determine j_o such that x_{j_o} is contained in $\{x_1, \ldots, x_m\}$ and $y_{k-1} < x_{j_o} < y_{k+1}$ and x_{j_o} maximizes $|\rho_k(x)|$.

Step 3.. If $|\rho_k(x_{j_o}| > |\rho_k(y_k)|$, replace y_k by x_{j_o}. Otherwise proceed to Step 4.

Step 4.. Set $k = k + 1$. If $k = n+2$ go to Step 5. Otherwise return to Step 2.

Step 5.. If n iterations occur without any change in $y_2, \ldots, y_{n+1}$, then $y_1, \ldots, y_{n+2}$ determine $P(x;\bar{a})$ and a halt occurs. Otherwise reset k to 2 and return to Step 2.

At first glance this algorithm appeared to be quite different from that given by Shenitzer and viewed strictly as a computational procedure it is different. However, it is possible in theory at least to reformulate Algorithm 3 so that the results obtained during each iteration for these two methods are identical. The principal difference

noted in the two algorithms is that in the latter the error function is maximized over each subinterval rather than the entire interval. But observing the results of Stiefel's Exchange Theorem (Theorem 9) and Equation (III-9) one sees that it is possible to modify Algorithm 3 to determine j_o in a subinterval. Then an exchange could be made in each subinterval similar to the procedure of Steps 2 and 3 of Algorithm 4.

Now, in the one case $|\rho_k(x)|$ is maximized over each subinterval and in the other case $|R(x)|$ is maximized over each subinterval. As was pointed out earlier, choosing j_o by $|R(x_{j_o})| = \max_{x \in S} |R(x)|$ does not necessarily yield the maximum increase in ρ. However, in the special case of Algorithm 3 each $|\alpha_i| = 1$ so that

$$|\rho^*| = \frac{\Sigma |\alpha_i| \, |\rho_i|}{\Sigma |\alpha_i|} = \frac{\Sigma |\rho_i|}{n+1} .$$

Hence choosing $\max |R(x)|$ yields the maximum increase in ρ. Thus the two algorithms would yield the same result during each iteration.

A most interesting and practical modification of the simplex algorithm (Algorithm 2) has been given by J. B. Kelly [18]. Given an error bound $\varepsilon > 0$ this algorithm determines the smallest integer n and a polynomial

$$P(x;\bar{a}) = a_o + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

of n^{th} degree, such that

$$|R(x)| = |f(x) - P(x;\bar{a})| < \varepsilon$$

for every $x \in S$.

The basis for the modification is the equation of Frobenius [19] which determines the inverse of a matrix A, when A is formed from another matrix B by the addition of one row and one column vector.

A revised algorithm will not be given. Instead, it will be shown how the equation of Frobenius could be used with Algorithm 2. Let $\varepsilon > 0$ be the required error bound and let $P(x;\bar{a})$ be the n_o^{th} degree polynomial obtained using Algorithm 2 so that

$$|f(x) - P(x;\bar{a})| \leq \rho_o .$$

for every $x \in S$.

If $\rho_o \leq \varepsilon$, $P(x;a)$ is the desired polynomial. Otherwise set $n = n_o + 1$ and add a new function, f_{n_o+1}, to the set $f_1, \ldots, f_{n_o}$ thus adding a new constraint to the linear program. Then the matrix of Step 1, Algorithm 2, is augmented by one row and one column vector and the inverse can be obtained by the equation of Frobenius. Actually this modification can also be used with Algorithm 1.

In the paper by Kelly n_o was initially set to zero and the best straight line fit was then obtained. This eliminated the necessity for a starting procedure.

While this modification is extremely practical since it determines automatically the degree of the polynomial necessary to satisfy the given error bound, it may result in more computations. This results for the choice of the starting reference set which is used each time the degree of the polynomial is increased; i.e., the new reference set

consists of the n points used to obtain the previous polynomial plus one additional point. This reference set would be quite different from the one suggested in (III-17). Thus it might be better to simply add a new point and restart the algorithm.

A few remarks concerning numerical results obtained using the various algorithms will be given next.

In the paper by Schenitzer [13] numerical results are given for the approximations of $\sin x$, $\cos x$ by a function of the form $\frac{P(x;\bar{a})}{g(x)}$, where $g(x) \neq 0$ over an internal S, and $P(x;\bar{a}) = a_0 + a_1 x + \cdots + a_n x^n$. The maximum error obtained was as high as 10^{-3} in some cases. Thus the algorithm did not appear to be suited for practical applications. This was also pointed out by H. H. Denman [20] in a paper on nearly optimized computer subroutines using Tchebysheff polynomials. However, in the more recent papers by Fraser and Hart [15] and Curtiss and Frank [17] excellent results were obtained using algorithems given above. Fraser and Hart used a variation of Algorithm 1 for the approximations $\sin x$, $\cos x$ over the intervals $(-0.3,0.3)$ and $(-1.3,1.3)$, respectively, using retional approximations of the form

$$\frac{a_0 + a_1 x^2 + a_2 x^4}{1 + b_2 x^2 + b_4 x^4}$$

for $\cos x$

$$\frac{a_1 x + a_3 x^3}{1 + b_2 x^2}$$

for sin x. The maximum errors obtained were 7.2×10^{-10} for cos x and 6.8×10^{-10} for sin x. Double precision arithmetic was used in both cases. The authors point out that when computing these coefficients using a computer of a given word size, it is necessary to use twice the word size in the computations in order to preserve single word length significance. Furthermore, the small interval over which the approximations were made can lead to an ill-conditioned matrix due to the differencing of small numbers. But this can also be overcome by the use of multiple precision arithmetic.

BIBLIOGRAPHY

1. Tchebysheff, P. L., Oeuvrer I, II, Petersbourg, (1899, 1907).

2. Clarkson, James A., Uniform Convex Spaces, Transactions American Mathematical Society, 40(1936), pp. 396-414.

3. Achieser, N. I., Theory of Approximation, Frederich Ungar, New York,(1956).

4. Haar, A., Die Minkowskische Geometrie und die Annäherung Stetige Funktionen, Math. Annalen, 78 (1918).

5. Bernstein, S. N., Lecons sur les proprietes extremaler et la meilleure approximations des functions analytique d'une variable reelle, Paris,(1926)

6. Buck, R. Creighton, Linear Spaces and Approximation Theory on Numerical Approximation, R. E. Langer, Editor, Madison, Wisconsin, (1959), pp. 11-23.

7. Mairhuber, John, On Haar's Theorem Concerning Tchebysheff Problems Having Unique Solutions, Proceedings American Mathematical Society, 7(1956), pp. 609-615.

8. Karlin, Samuel, Mathematical Methods and Theory in Games, Programming and Economics, Addison-Wesley, Reading, Massachusetts, 1(1959).

9. Stiefel, E. L., Numerical Methods of Tchebysheff Approximations, R. E. Langer, Editor, Madison, Wisconsin, (1959), pp. 217-232.

10. Stiefel, D. L., Über diskrete und lineare Tchebysheff Approximation, Numerische Mathematik, 1(1959), pp. 1-28.

11. Stiefel, E. L., Note on Jordan Elimination, Linear Programming and Tchebysheff Approximation, Numerische Mathematik, 1(1959), pp. 253-268.

12. Novodvarskii, E. N., and I. Sh. Pinsker, On a Process of Equalization of Maxima, Usp. Mat. Nauk., 6(1951), pp. 174-181 (English translation by A. Shenitzer available from New York University).

13. Shenitzer, A., Chebychev Approximation of a Continuous Function by a Class of Functions, Journal of Association for Computing Machinery, 4(1957), pp. 30-35.

14. Cheney, E. W., and A. A. Goldstein, Newton's Method for Convex Programming and Tchebysheff Approximation, Numerische Mathematik, 1(1959), pp. 253-268.

15. Fraser, W., and J. F. Hart, On the Computation of Rational Approximation to Continuous Functions, Communications of the Association for Computing Machinery, Vol. 5, No. 7,(1962), pp. 401-403.

16. Gass, Saul I., Linear Programming, McGraw-Hill, New York (1958).

17. Curtiss, P. C., Jr., and W. L. Frank, An Algorithm for the Determination of the Polynomial of Best Minimax Approximation to a Function Defined on a Finite Point Set, Journal of the Association for Computing Machinery, Vol. 6, No. 3, (1959), pp. 395-404.

18. Kelly, James E., Jr., An Application of Linear Programming to Curve Fitting, Journal of the Society of Industrial and Applied Mathematics, Vol. 6, No. 1 (1958), pp. 15-22.

19. Bodewig, E., Matrix Calculus, North Holland Publishers, Amsterdam, (1956).

20. Denman, H. H., Computer Generation of Optimized Subroutines, Journal of the Association for Computing Machinery, Vol. 8, No. 1 (1961).

21. C. de la Vallee-Poussim, Lecons sur l'approximation des fonctions d'une variable reele, Gauthier-Villars, Paris, (1952).