SEG-TR-66-32

A METHOD FOR CALCULATING POLYNOMIALS OF BEST APPROXIMATION

ALFRED OLIVER

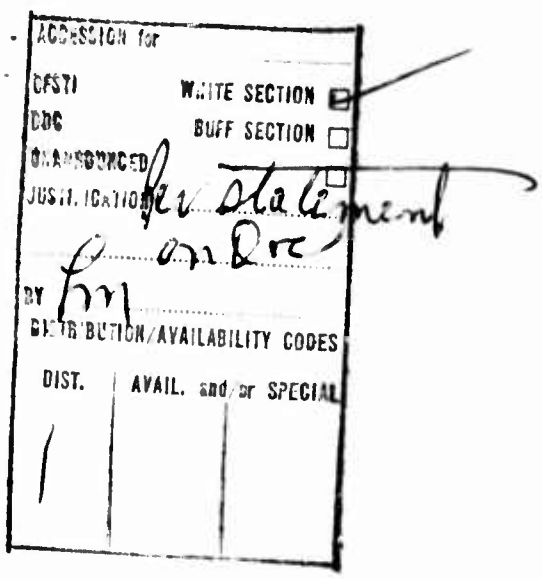
TECHNICAL REPORT SEG-TR-66-32

AUGUST 1966

SYSTEMS ENGINEERING GROUP
RESEARCH AND TECHNOLOGY DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.



Copies of this report should not be returned to the Research and Technology Division unless return is required by security considerations, contractual obligations, or notice on a specific document.

100 - October 1966 9-204

SEG-TR-66-32

A METHOD FOR CALCULATING POLYNOMIALS OF BEST APPROXIMATION

ALFRED OLIVER

FOREWORD

This report was prepared in the Digital Computation Division, Research and Technology Division, during the period from 1 June to 1 September 1964. It was separately submitted as a thesis in partial fulfillment of the requirements for the degree of Master of Science from The Ohio State University.

The author wishes to acknowledge the very significant contribution made by Dr. Theodore W. Hildebrandt of The Ohio State University.

This report was submitted by the author as a technical report in June 1966.

This technical report has been reviewed and is approved.

DEMETRIUS C. ZONARS
Digital Computation Division
Directorate of Computation

ABSTRACT

This report deals with minimax approximations to functions defined on the real line. These approximations are of particular use to mathematicians having access to a digital computer facility.

The cases considered are for functions defined as discrete points on the Cartesian plane and for continuous, bounded functions defined on a closed interval of the real line. A method for obtaining the approximations using linear programming techniques is presented for each case.

An analysis of the error function and a brief outline of the computer program are included.

TABLE OF CONTENTS

SECTION I

INTRODUCTION

The general problem considered in this paper is the approximation of a function f by an approximating function F. There are two main reasons for being interested in approximating a given function. In the first place, if we are using a digital computer, nonelementary functions may take considerable machine time to compute, and if we are to evaluate such a function a large number of times, it will be advantageous to replace it with a more elementary one. In the second place, if the function f is given in tabular form, a good approximating function will replace the need for table look-up and interpolation and will economize on storage.

In approximating a given function we are faced with two major problems: (1) to determine the form of the approximating function and (2) to determine a criterion for the accuracy of the approximation. The primary consideration in determining the type of approximating function to use is that this function be compatible with the function to be approximated. Thus we would not approximate a function which has a singularity with a polynomial. There may be, however, certain practical considerations which influence the choice of the approximating function. In the following we limit ourselves to polynomials and functions which we can approximate with polynomials, since they are easily evaluated.

Just as there are numerous choices for the approximating function, there are many ways of measuring the accuracy of the approximation. Among the more common methods are exact matching at the sample points; least-squares, which is well suited to data subject to a random error; and Chebycheff, which reduces the maximum error to a minimum. The actual choice of the criterion to use depends on the particular problem. The Chebycheff criterion is particularly useful when the noise connected with the data is negligible, or when it is desired to keep the maximum error minimal. We will concern ourselves with the Chebycheff criterion.

The problem we will consider is that of finding, among all polynomials Q of degree less than or equal to n, a polynomial P such that given a function f defined on a set S of real numbers

$$\max_{x \in S} | f(x) - P(x) | = \min_{Q} (\max_{x \in S} | f(x) - Q(x) |).$$

The set S may be a discrete set of points or an interval on the real line. P is said to be the polynomial which best approximates f on S in the Chebycheff sense.

We will show the existence and uniqueness of such a polynomial P, describe the characteristics of the error function, and develop the necessary computational methods to find P.

SECTION II

CHEBYCHEFF APPROXIMATION TO A FUNCTION DEFINED ON A FINITE POINT SET

Given a function f defined on a point set S of m points of the real line, we want to find a polynomial P, of degree less than or equal to n, $n+2 \leq m$, for which the error

$$\max_{x \in S} | f(x) - P(x) |$$

is a minimum. We call P the polynomial of degree less than or equal to n of best approximation to f on S.

The following theorem of C. de la Vallee Poussin (Reference 19) characterizes the polynomial P.

Theorem: Let P be a polynomial of degree less than or equal to n, of best approximation to a function f defined on a point set S of $m \geq n+2$ points.

Then P is uniquely characterized by the existence of at least n+2 points of S at which the difference $|f(x) - P(x)|$ assumes its maximum value on S, with the algebraic sign of $f(x) - P(x)$ alternating at these points.

Proof: We suppose $|f(x) - P(x)|$ assumes its maximum value at no more than n+1 points of the set S and arrive at a contradiction.

Let S' be this set of at most n+1 points. We can then find a polynomial Q of degree less than or equal to n such that Q(x) has the same algebraic sign as $f(x) - P(x)$ for every point x in S'.

Let $$E(P;S) = \max_{x \in S} |f(x) - P(x)|$$

$$E(P;S-S') = \max_{x \in S-S'} |f(x) - P(x)|$$

and $$|| Q(S) || = \max_{x \in S} |Q(x)| .$$

Since $E(P;S) > E(P;S-S')$ we can find an $\epsilon > 0$ such that $\epsilon || Q(S) || < E(P;S) - E(P;S-S')$.

Let $R(x) = P(x) + \epsilon Q(x)$. R(x) is then a polynomial of degree less than or equal to n. We now show that under the hypothesis that there are at most n+1 points at which $|f(x) - P(x)|$ assumes its maximum value on S, the polynomial R(x) is a better approximation to f(x) than P(x), which is a contradiction since P(x) was taken as the best approximation to f(x) on S.

On the subset S' of S we have

$$|f(x) - R(x)| = |f(x) - P(x) - \epsilon Q(x)| = |f(x) - P(x)| - \epsilon |Q(x)|.$$

The last equality holds because Q(x) has the same algebraic sign as $f(x) - P(x)$ on S'.

But $|f(x) - P(x)| - \epsilon |Q(x)| < |f(x) - P(x)|$. Hence for $x \in S'$ we have $|f(x) - R(x)| < |f(x) - P(x)|$.

On the subset S-S' of S we have

$$|f(x) - R(x)| = |f(x) - P(x) - \epsilon Q(x)| \leq |f(x) - P(x)| + \epsilon ||Q(S)|| .$$

Then

$$\max_{x \epsilon S-S'} |f(x) - R(x)| \leq \max_{x \epsilon S-S'} |f(x) - P(x)| + \epsilon ||Q(S)|| .$$

But

$$\max_{x \epsilon S-S'} |f(x) - P(x)| + \epsilon ||Q(S)|| < E(P;S-S') + (E(P;S) - E(P;S-S'))$$

$$< E(P;S)$$

Hence $\max_{x \epsilon S-S'} |f(x) - R(x)| < E(P;S)$.

Since S is the union of S' and (S-S') the polynomial R(x) gives a better approximation to f(x) on S than P(x), which contradicts the hypothesis that P(x) is the polynomial of best approximation.

Having characterized the polynomial P of best approximation, we now show it is unique.

Theorem: Let P be a polynomial of degree less than or equal to n, of best approximation to a function f defined on a set S of $m \geq n+2$ points. Then P is unique.

Proof: Assume the contrary, that there exists a polynomial Q of degree less than or equal to n, different from P, such that $\max_{x \epsilon S} |f(x) - P(x)| = \max_{x \epsilon S} |f(x) - Q(x)|$.

Let $E(S) = \max_{x \epsilon S} |f(x) - P(x)| = \max_{x \epsilon S} |f(x) - Q(x)|$.

Then for $x \epsilon S$

$$-E(S) \leq f(x) - P(x) \leq E(S)$$

and $$-E(S) \leq f(x) - Q(x) \leq E(S).$$

Addition and division by two yields

$$-E(S) \leq f(x) - .5(P(x) + Q(x)) \leq E(S).$$

Then $R(x) = .5(P(x) - Q(x))$ is also a polynomial of degree less than or equal to n of best approximation to f(x) on S.

Consider a point x' of S for which $f(x') - R(x') = E(S)$.
Then $2(f(x') - R(x')) = (f(x') - P(x')) + (f(x') - Q(x')) = 2E(S)$.
Since both $f(x') - P(x')$ and $f(x') - Q(x')$ are less than or equal to E(S) we must have

$f(x') - P(x') = f(x') - Q(x')$, i.e., $P(x') = Q(x')$.

It can likewise be shown that for any point x" of S for which f(x") - R(x") = -E(S) we will have P(x") = Q(x").

By the previous theorem there are at least n+2 points at which f(x) - R(x) = ±E(S). Hence P(x) and Q(x) coincide at a minimum of m+2 points of S. But then P(x) and Q(x) are identical since two polynomials of degree less than or equal to n, coinciding at n+2 points, coincide everywhere.

We now return to the problem of finding the unique polynomial P, of degree less than or equal to n, of best approximation to a function f defined on a set S of m points, where m is greater than or equal to n+2.

The problem is one of determining the coefficients $c_0, \ldots, c_n$ and the deviation E(S) such that

$$E(S) = \max_{x \in S} \left| f(x) - \sum_{j=0}^{j=n} c_j x^j \right|$$

is a minimum.

Let $a_{ij} = x_i^j$, $1 \le i \le m$ and $0 \le j \le n$. The problem is then to find $c_0, c_1, \ldots, c_n$ and E(S) such that

$$E(S) + \sum_{j=0}^{n} a_{ij} c_j \ge f_i \qquad i = 1, \cdots, m$$

$$E(S) - \sum_{j=0}^{n} a_{ij} c_j \ge -f_i$$

$$E(S) \ge 0$$

and $E(S)$ is a minimum.

We now have a system of 2m+1 linear inequalities in the n+2 unknowns $c_0, c_1, \ldots, c_n$ and E(S), and a linear functional Z = E(S) to be minimized.

We will see that there is an efficient way of adapting our approximation problem to a linear programming form. To show this we define certain terms and concepts of linear programming.

Associated with every linear programming problem is another linear programming problem called the 'dual'. Below is given the relation between a typical 'primal' problem and its dual. It should be noted that either of the two problems may be called the primal, and the other would then be the dual.

Primal Problem: Find a column vector $X = (x_1, \ldots, x_q)$ which maximizes the linear functional

$$Z = CX$$

subject to the restrictions

$$AX = B \text{ and } X \ge 0$$

where $C = (c_1, \ldots, c_q)$ is a row vector, $B = (b_1, \ldots, b_p)$ is a column vector, and $A = (a_{ij})$ is a p by q matrix.

Dual Problem: Find a row vector $Y = (y_1, \ldots, y_p)$ which minimizes the linear functional

$$W = YB$$

subject to the restriction $YA \geq C$, where A, B, and C are defined as in the primal problem. Note that there is no restriction on the sign of Y.

The following is the well-known duality theorem, which establishes the important relation between a primal problem and its dual.

We define a feasible solution as one which satisfies the constraints of the given problem, and an optimal solution as a feasible solution which maximizes or minimizes the object function Z or W, as the case may be.

Theorem: If either the dual or the primal problem has a finite optimal solution, the other problem has a finite optimal solution, and the extremes of the linear functionals are equal, i.e., max Z = min W.

A proof of this theorem may be found in the book, Linear Programming, by S. I. Gass (Reference 7).

We solve the above linear programming problem using the Simplex Method due to G. B. Dantzig (Reference 5). A 'basic' feasible solution is one with, at most, p positive variables, and one which expresses B in terms of p linearly independent vectors P_j, $j = 1, \ldots, p$, where $P_j = (a_{1j}, \ldots, a_{pj})$ is the jth column of A.

The algorithm starts with a basic feasible solution. We can always find such a solution, though it may be necessary to introduce certain artificial variables to do so. If this solution is not optimal, we can either find a class of solutions in which there is no bound on the object function or we can construct a new basic feasible solution in which the value of the object function is not less than that for the previous solution. The simplex algorithm consists of repeating this cycle until we have constructed an optimal basic feasible solution or a class of feasible solutions for which the object function is unbounded.

Let G be the p by p matrix formed by the p linearly independent vectors P_j from A which correspond to the optimal feasible solution. We call G the optimal basis. Let X' be the p dimensional optimal basic feasible solution. At the termination of the simplex algorithm there will be a p dimensional row vector C' such that

$$X' = G^{-1}B \text{ and } \max Z = C'X'$$

If we let Y' be the optimal basic feasible solution of the dual, then

$$\min W = Y'B = C'G^{-1}B = C'X' = \max Z, \text{ i.e., } Y' = C'G^{-1}.$$

Most standard linear programming codes for digital computers will yield the solution to both the primal and the dual problems.

Consider now our approximation problem. Since $a_{io} = x_i^o = 1$ we let the matrix A be

$$\begin{pmatrix} 1 & 1 & \cdots & 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & \cdots & 1 & -1 & \cdots & -1 & 0 \\ a_{11} & a_{21} & \cdots & a_{m1} & -a_{11} & \cdots & -a_{m1} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} & -a_{1n} & \cdots & -a_{mn} & 0 \end{pmatrix}$$

We also let

$$F = (f_1, \ldots, f_m, -f_1, \ldots, -f_m, 0)$$

$$Y = (E(S), c_0, \ldots, c_n)$$

and

$$U = \begin{pmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix}$$

where U is a 2m+1 by 1 column vector. The problem is then to find a row vector Y which minimizes the linear functional $W = YU$ subject to $YA \geq F$, which is precisely the dual form described above. We can consider the primal problem as one of finding a column vector X which maximizes $Z = FX$ subject to $AX = U$ and $X \geq 0$.

We can now solve this primal version of the problem using a standard linear programming code, which will yield a solution to the dual as well.

There are three reasons for taking the "dual approach" rather than solving for Y directly:

1. Since Y is not restricted in sign, we would be forced to add variables to our problem before attempting a solution. The dual approach eliminates this need.

2. The simplex method is most efficient in terms of time when the number of variables is greater than the number of equations. Since in our problem we have n+2 variables but 2m+1 equations, where $m \geq n+2$, it is advantageous to consider the dual so that we have 2m+1 variables and n+2 equations.

3. The addition of points at which we want to approximate f(x) corresponds to the addition of variables in the problem as we consider it, and this can be done without restarting the algorithm at its initial phase. But addition of points would mean the addition of constraints if we solve our problem directly. Having obtained a solution vector which satisfies the original constraints, we would have to start the solution anew to take into consideration the new constraints.

Interpolation Error: Let f be a continuous function with at least n+2 derivatives, defined on a set S of m points. Having obtained a polynomial P, of degree less than or equal to n, $m \geq n+2$, which is a best approximation to f(x), we may want to use P as an interpolating polynomial to interpolate between the given points of S.

We will call the points of S at which $|f(x) - P(x)|$ takes on its maximum value the critical points, and we assume that there are exactly n+2 of them. We further assume that the smallest and largest of the points in S are in the critical set. Let x' and x" be these two points. Given these conditions, we can obtain an interpolation error for P over $[x', x'']$.

Let $e(i) = f(x_i) - P(x_i)$, $i = 1, \ldots, n+2$, for x_i in the critical set. We can construct a polynomial L(x) of degree n+1 which will assume the values e(i) at the points x_i of the critical set. We use the Lagrangian interpolation formula to obtain

$$L(x) = h(1, x)\, e(1) + \ldots + h(n+2, x)\, e(n+2)$$

where

$$h(i,x) = \frac{(x - x_1) \ldots (x - x_{i-1}) (x - x_{i+1}) \ldots (x - x_{n+2})}{(x_i - x_1) \ldots (x_i - x_{i-1}) \quad (x_i - x_i - x_{i+1}) \ldots (x_i - x_{n+2})}$$

The error term of this formula is

$$R(x) = \frac{G^{(n+2)}(v)\, Q(x)}{(n+2)}$$

where $Q(x) = (x - x_1) \ldots (x - x_{n+2})$.

The x_i's are the critical points, and $G(x) = f(x) - P(x)$, and the point v is some point in $[x', x'']$. A complete discussion of the Lagrangian interpolation formula and the error term can be found in (Reference 8).

Since P is a polynomial of degree at most n, its (n+2)nd derivative is identically zero; hence,

$$G^{(n+2)}(v) = f^{(n+2)}(v) \text{ and } R(x) = \frac{f^{(n+2)}(v)\, Q(x)}{(n+2)!}$$

The error f(x) - P(x) is then equal to L(x) + R(x), and this is the error incurred when using P(x) as an interpolating polynomial.

SECTION III

CHEBYCHEFF APPROXIMATION TO A FUNCTION DEFINED ON A CLOSED INTERVAL ON THE REAL LINE

Let f be a given function defined and continuous on a closed interval $[a, b]$. We are concerned with finding a polynomial P, of degree less than or equal to n, such that the error

$$\max_{a \le x \le b} |f(x) - P(x)|$$

is a minimum.

The following theorem of Chebycheff (Reference 1) characterizes the polynomial P.

Theorem: Among all the polynomials of degree less than or equal to n defined on the real line, the polynomial P which deviates least from a given continuous function is uniquely characterized by the fact that in the interval $[a, b]$ over which we approximate the given function, the number of consecutive points at which the difference $f(x) - P(x)$, with alternating signs, assumes the value

$$\max_{a \le x \le b} |f(x) - P(x)|$$

is not less than n+2.

The next theorem is important for computational reasons. We define a local maximum as follows: we say f has a local maximum at a point x if there exists a $\delta > 0$ such that $f(y) \le f(x)$ for all y such that $0 < |x - y| < \delta$. A local minimum can be defined similarly by reversing the inequality between f(x) and f(y). We now show that the function

$$E(P; [a, b]) = |\max_{a \le x \le b} f(x) - P(x)|,$$

where $P(x) = c_0 + c_1 x + \ldots = c_n x^n$ is an arbitrary polynomial of degree $\le n$, is convex, and hence a local maximum or minimum is the unique maximum or minimum for the whole interval $[a, b]$.

Theorem: Let θ be in $[0, 1]$, and let P_1 and P_2 be polynomials of degree less than or equal to n. Then

$$E(\theta P_1 + (1 - \theta) P_2; [a, b]) \le \theta E(P_1; [a, b]) + (1 - \theta) E(P_2; [a, b])$$

Proof:

$$E(\theta P_1 + (1-\theta)P_2; [a,b]) = \max_{a\le x\le b} |f(x) - \sum_{i=0}^{n} (\theta c_{1_i} + (1-\theta) c_{2_i}) x^i |$$

$$= \max_{a\le x\le b} |\theta (f(x) - \sum_{i=0}^{n} c_{1_i} x^i) + (1-\theta)(f(x) - \sum_{i=0}^{n} c_{2_i} x^i)|$$

$$\le \max_{a\le x\le b} (\theta |f(x) - \sum_{i=0}^{n} c_{1_i} x^i| + (1-\theta) |f(x) - \sum_{i=0}^{n} c_{2_i} x^i|)$$

$$\le \theta \max_{a\le x\le b} |f(x) - \sum_{i=0}^{n} c_{1_i} x^i| + (1-\theta) \max_{a\le x\le b} |f(x) - \sum_{i=0}^{n} c_{2_i} x^i|$$

$$= \theta E(P_1; [a,b]) + (1-\theta) E(P_2; [a,b]).$$

Let us suppose now that the function E has local minima for polynomials P_1 and P_2, of degree less than or equal to n, and let θ be in $[0, 1]$. We have just proved that

$$E(\theta P_1 + (1-\theta) P_2; [a, b]) \le \theta E(P_1; [a, b]) + (1-\theta) E(P_2; [a, b]).$$

Without loss of generality, suppose $E(P_1; [a, b]) < E(P_2; [a, b])$. Then by letting θ approach 0 we can find a P_3 as near to P_2 as we like such that $E(P_3; [a, b]) < E(P_2; [a, b])$, contradicting the hypothesis that E has a local minimum to P_2. By P_3 being near to P_2 we mean that

$$(\sum_{i=0}^{n} (c_{3_i} - c_{2_i})^2)^{1/2} < \epsilon, \quad i = 0 \cdots, n,$$

for arbitrary $\epsilon > 0$.

On the other hand, if $E(P_1; [a, b]) = E(P_2; [a, b])$, then there can be no P_3 such that

$$c_{1_i} < c_{3_i} < c_{2_i} \text{ or } c_{1_i} > c_{3_i} > c_{2_i}, \; i = 0, \ldots, n,$$

and such that $E(P_3; [a, b]) > E(P_1; [a, b])$ and $E(P_3; [a, b]) > E(P_2; [a, b])$. We can see this by observing that $E(\theta P_1 + (1-\theta) P_2; [a, b]) \le E(P_1; [a, b])$.

We see then that the function E has no more than one minimal value; and hence, once we have a minimum for E, we know this is the unique minimal value of the function.

We now develop an algorithm for determining approximately the unique polynomial P of degree less than or equal to n which is a best approximation to a given function f, defined and continuous on an interval $[a, b]$. Without loss of generality, we can make the linear substitution $x' = (2x - a - b) / (b - a)$ which transforms the interval $[a, b]$ into the interval $[-1, 1]$.

The exact solution to the problem involves the solution of a system of nonlinear equations, and is quite difficult to arrive at in all but the simplest cases. We can, however, arrive at an approximate solution by successive approximations on discrete subsets of the interval $[-1, 1]$.

We generate a sequence $Q(k; x)$ of polynomials of degree less than or equal to n, along with a sequence

$$X(k) = (x(k, i): i = 1, 2, \ldots, n+2)$$

of sets of n+2 points of the interval $[-1, 1]$. The polynomial $Q(k; x)$ is the best approximation to $f(x)$ on $X(k)$, $k = 0, 1, 2, \ldots$

Let

$$E(Q(k; x); X(k)) = \max_{x \in X(k)} |Q(k; x) - f(x)|.$$

Any set $X(0)$ of n+2 points from $[-1, 1]$ for which $E(Q(0; x); X(0))$ is not equal to zero can be chosen to start the calculations, but in practice it has been found that faster convergence is attained if we choose the n+2 points at which the Chebycheff polynomial $T(n+1)$ of degree n+1 attains its extremal values.

Having chosen $X(0)$, we find $Q(0; x)$ by the method described in Section II. The iterative method proceeds as follows:

1. Given $Q(k; x)$, choose

$X(k + 1) = (x(k + 1, i): x(k + 1, i) < x(k + 1, i + 1))$, $i = 0, \ldots, n+1$, so that $Q(k; x) - f(x)$ has opposite signs at adjacent points of $X(k+1)$ and $Q(k; x) - f(x)$ assumes its extrema on $X(k+1)$. It should be noted that $X(k + 1)$ and $X(k)$ need not be disjoint. It will be seen that for most cases the two sets do indeed have elements in common.

Let

$$M(k) = \max_{x \in X(k + 1)} |Q(k; x) - f(x)|$$

and

$$m(k) = \min_{x \in X(k + 1)} |Q(k; x) - f(x)|$$

We shall prove that $m(k)$ will be at least equal to $E(Q(k; x); X(k))$.

2. $Q(k + 1; x)$ and $E(Q(k + 1; x); X(k + 1)$ are then determined so that $Q(k + 1, x)$ is the polynomial of degree less than or equal to n of best approximation to $f(x)$ on $X(k + 1)$.

We accomplish Steps 1 and 2 in the following manner. Since the error $Q(k; x) - f(x)$ takes on the values $E(Q(k; x); X(k))$ and $-E(Q(k; x); X(k))$ at no less than n+2 points of $X(k)$, there

will be at least n+1 points of $[-1, 1]$ for which $Q(k; x) - f(x)$ is equal to zero, say $x_1' < x_2' \ldots < x_{n+1}'$. Let $x_0' = -1$ and $x_{n+2}' = 1$

where $x_0' < x_1$ and $x_{n+1}' < x_{n+2}'$.

The set $X(k+1)$ is then chosen so that $x(k+1, i)$ is the point on $[x_{i-1}', x_i']$ at which $|Q(k; x) - f(x)|$ is maximal. The set $X(k+1)$ so chosen satisfies the requirements in Step 1. We now compute $Q(k+1; x)$ and $E(Q(k+1; x); X(k+1))$.

Let

$$d(i) = |Q(k; x(k+1, i) - f(x(k+1, i)|, \; i = 1, \ldots, n+2.$$

Then

$$(Q(k; x(k+1, i)) - f(x(k+1; i))) - (Q(k+1; x(k+1, i)) - f(x(k+1, i))) =$$

$$Q(k; x(k+1, i)) - Q(k+1; x(k+1, i)) = \pm(d(i) - E(Q(k+1; x(k+1, i)))).$$

We see that $E(Q(k+1; x); X(k+1))$ must be greater than $m(k)$, for if $m(k) \geq E(Q(k+1; x); X(k+1))$ the polynomial $Q(k; x) - Q(k+1; x)$ would have n+2 zeroes; but since $Q(k; x)$ and $Q(k+1; x)$ are both of degree less than or equal to n, this would imply that $Q(k; x) \equiv Q(k+1; x)$.

We can iterate the procedure so that we obtain

$$E(Q(k; x); X(k)) \leq m(k) < E(Q(k+1; x); X(k+1)), \; k = 0, 1, 2, \ldots .$$

Let $P(x)$ be the polynomial of best approximation to $f(x)$ on $[-1, 1]$, and let

$$E(P; [-1, 1]) = \max_{-1 \leq x \leq 1} |f(x) - P(x)|$$

For $Q(k+1; x)$ not identical to $P(x)$, the following relation holds:

$$E(Q(k+1); X(k+1)) < E(P; [-1, 1]) < M(k)$$

Since $X(k+1)$ is a subset of $[-1, 1]$, and $Q(k+1)$ is the polynomial of best approximation to $f(x)$ on $X(k+1)$, we cannot have $E(Q(k+1); X(k+1)) \geq E(P; [-1, 1])$, for this would mean $P(x)$ is a better approximation to $f(x)$ on $X(k+1)$, which is a contradiction.

It is also seen that we cannot have $M(k) \leq E(P; [-1, 1])$ since if we did then $Q(k+1)$ would be a better approximation to $f(x)$ on $[-1, 1]$ than $P(x)$, which is again a contradiction.

We now show that the process described above will indeed result in convergence of the sequence $E(Q(k; x); X(k))$, $k = 0, 1, 2, \ldots$ to the number $E(P; [-1, 1])$.

We can find a λ such that $0 < \lambda < 1$ and

1. $E(Q(k+1; x); X(k+1)) - m(k) > (1 - \lambda)(M(k) - m(k))$

Since $M(k) > E(P; [-1, 1]) > m(k)$, we also have

2. $E(Q(k+1; x); X(k+1)) - m(k) > (1 - \lambda)(E(P; [-1, 1]) - m(k))$

and $m(k+1) \geq E(Q(k+1; x); X(k+1))$ further implies

3. $m(k+1) - m(k) > (1-\lambda)\,(E(P; \left[-1, 1\right]) - m(k))$

Subtracting both sides of (3) from $E(P; \left[-1, 1\right]) - m(k)$, we obtain

4. $E(P; \left[-1, 1\right]) - m(k+1) < \lambda\,(E(P; \left[-1, 1\right]) - m(k))$

But 4. implies that the quantity $E(P; \left[-1, 1\right]) - m(k)$, $k = 0, 1, 2, \ldots$, will tend to zero at least as fast as the terms of a diminishing geometric progression upon repeated applications of the algorithm. Since $E(Q(k+1; x); X(k+1)) > m(k)$, the quantity $E(P; \left[-1, 1\right]) - E(Q(k+1; x); X(k+1))$ will likewise tend to zero at the same rate.

Note also that by 1. we have

$$M(k) - E(P; \left[-1, 1\right]) \quad M(k) - m(k) < \frac{E(Q(k+1; x); X(k+1)) - m(k)}{(1-\lambda)}$$

$$< \frac{E(P; \left[-1, 1\right]) - m(k)}{(1-\lambda)}$$

i.e.,

$$M(k) - E(P; \left[-1, 1\right]) < \frac{E(P; \left[-1, 1\right]) - m(k);}{(1-\lambda)}$$

therefore, $\lim_{k \to \infty} M(k) = E(P; \left[-1, 1\right])$.

The search for the set $X(k+1)$ having obtained $X(k)$ is done in the following manner: Start at a point $x(k, i)$ of $X(k)$ and using a step size h we find the direction of increasing $|Q(k; x) - f(x)|$ and proceed in this direction. The search continues until there are two points straddling the required peak of the curve $Q(k; x) - f(x)$. The point $x(k+1; i)$ is then chosen as the extreme of a quadratic interpolation which uses the last three points. If the end points -1 and 1 are included in $X(k)$, we restrict the search at these points to one direction only, i.e., to the inside of $[-1, 1]$.

There is an important result of a paper by Murnaghan and Wrench (Reference 18) which relates to this search for a set $X(k+1)$. They found that the partial derivative of a coefficient of $P(x)$ with respect to an interior point of the critical set corresponding to $P(x)$ is zero, and, therefore, this optimal critical set need not be determined with extreme accuracy. This result is valid only if $f(x)$ is differentiable as well as continuous.

SECTION IV

UPPER BOUND FOR THE MINIMAX ERROR

Let P(x) be the polynomial of degree less than or equal to n of best approximation to a continuous function f(x) defined on $\left[-1, 1\right]$.

Let

$$E' = \max \mid f(x) - P(x) \mid \text{ for } x \text{ in } \left[-1, 1\right].$$

We will call E' the minimax error for P(x) on $\left[-1, 1\right]$.

Let f(x) have a Fourier-Chebycheff expansion

$$T(x) = 0.5\, a_0 + \sum_{k=1}^{\infty} a_k T_k(x)$$

where $T_k(x)$ is the Chebycheff polynomial of degree k.

Theorem (2): Let $\left\{a_{i_k}\right\}$ be the subsequence of all non-zero coefficients of T(x). If for all k >n we have

$$\mid a_{i_{k+1}} / a_{i_k} \mid \leq \lambda < 1$$

then for k>n we have $E' \leq \mid a_{i_{n+1}} \mid \frac{1}{1-\lambda}$, where n is the degree of P(x).

Proof: If for all k>n we have $\mid a_{i_{k+1}} / a_{i_k} \mid \leq \lambda < 1$ then $\sum_{k=1}^{\infty} \mid a_k \mid < \infty$. Hence for k>n

$$E' \leq \sum_{j=1_{n+1}}^{\infty} a_j T_j(x) \leq \sum_{j=1_{n+1}}^{\infty} \mid a_j \mid,$$

since max $T_j(x) = 1$ on $\left[-1, 1\right]$.

But

$$\sum_{j=i_{n+1}}^{\infty} \mid a \mid \leq \mid a_{i_{n+1}} \mid (1 + \lambda + \lambda^2 + \cdots) = \mid a_{i_{n+1}} \mid \frac{1}{1-\lambda}.$$

This completes the proof.

SECTION V

NUMERICAL RESULTS

All calculations were performed on an IBM 7094 Digital Computer. Twenty-seven bit accuracy was used throughout. The linear programming code is that of R. J. Clasen of the RAND Corporation. The subroutine name is RSMSUB, with SHARE Distribution Number 1281.

For most functions it was found that the sample independent variable set need not contain more than n+2 points (where n is the degree of the approximating polynomial), but in some cases it was found advantageous to use a larger sample as this shortened the search for the critical set considerably. For all the cases considered the end points of the interval over which the approximation was obtained were included in the critical set.

We can in fact show that if f(x) has at least (n + 1) derivatives and if $f^{(n+1)}(x)$ does not change signs in $[a, b]$, then the end points a and b will always be included in the critical set.

Let $X = \{x_1, x_2, \ldots, x_{n+2}\}$ be the critical set of sample points. Let D(x) = f(x) - P(x), where f(x) is the function we are to approximate, and P(x) is the polynomial of degree less than or equal to n which best approximates f(x) in the interval $[a, b]$. Let $P(x) = c_0 + c_1 x + \ldots + c_n x^n$.

We know that at the interior points of X the derivative of D(x) must be zero, i.e.,

$$D'(x_i) = 0 \qquad i = 2, 3, \ldots, n+1$$

In order to be in the set X, the points x_1 and x_{n+2} must either be end points of the interval, so that $x_1 = a$ and $x_{n+2} = b$, or must be such that $D'(x_1) = D'(x_{n+2}) = 0$. We show that given the above conditions the latter cannot be true.

We know that D'(x) is equal to zero at least n times in $[a, b]$. Repeated applications of Rolle's theorem show that $D^{(n)}(x)$ is equal to zero at least once in $[a, b]$. Since P(x) is of degree at most n we have $D^{(n+1)}(x) = f^{(n+1)}(x)$. Hence, if $f^{(n+1)}(x)$ does not change sign in $[a, b]$, neither will $D^{(n+1)}(x)$. But then $D^{(n)}(x)$ is either monotonic increasing or decreasing, and hence must equal zero exactly once. Then $D^{(n-1)}(x)$ will be zero at most twice in $[a, b]$, and continuing in this way, we see that D'(x) will equal zero at most n times, and, by what was said above, exactly n times. Then D'(x) cannot be equal to zero for x_1 and x_{n+2}; hence, we must have $x_1 = a$ and $x_{n+2} = b$.

This result is useful since if f(x) satisfies the required conditions on $[a, b]$, then we can include a and b in our critical set from the start.

Example 1: We compute the polynomial P(x) of degree less than or equal to 3 which is the best approximation to a function f(x) given as a set of ordered pairs $(x_i, f(x_i))$. We use as a sample set X, 11 equally spaced points ranging from 0 to 1. The values $f(x_i)$ are taken to be the values of the function sin(x) on the sample set X. The Fortran Library subroutine was used to obtain the values $f_i = f(x_i) = \sin(x_i)$.

$$P(x) = c_0 + c_1x + c_2x^2 + c_3x^3$$

$c_0 = -0.0001472$ $\qquad$ $c_2 = -0.0193834$

$c_1 = 1.00444$ $\qquad$ $c_3 = -0.143585$

$$D = \max_{1 \le i \le 11} \left| f(x_i) - P(x_i) \right| = 0.0001472$$

x_i	$f(x_i)$	$P(x_i)$	$E(x_i) = f(x_i) - P(x_i)$
0.0	0.0	-0.0001472	0.0001472
0.1	0.0998334	0.0999592	-0.0001258
0.2	0.198669	0.198817	-0.0001472
0.3	0.295520	0.295563	-0.0000429
0.4	0.389418	0.389338	0.0000808
0.5	0.479426	0.479278	0.0001472
0.6	0.564642	0.564524	0.0001187
0.7	0.644218	0.644213	0.0000051
0.8	0.717356	0.717483	-0.0001271
0.9	0.783327	0.784474	-0.0001472
1.0	0.841471	0.841324	0.0001472

Example 2

Function	$1/\sqrt{2\pi} \int_{-\infty}^{x} \exp(-0.5t^2)\,dt$
Range	$[4, 4]$
Approximation P(x)	$c_0 + c_1x + c_2x^2 + c_3x^3$
Number of Iterations	3
E(Q(2; x); X(2))	0.0645903
c_0	0.5
c_1	0.288445
c_2	0.0
c_3	-0.011225

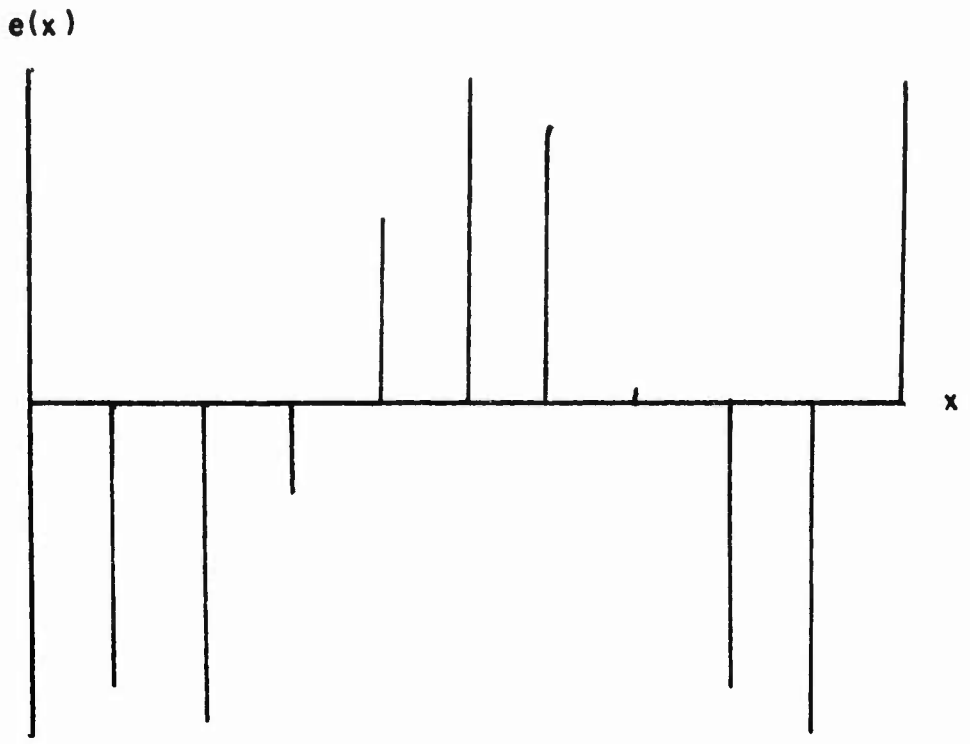


Figure 1. Error Function for Example 1

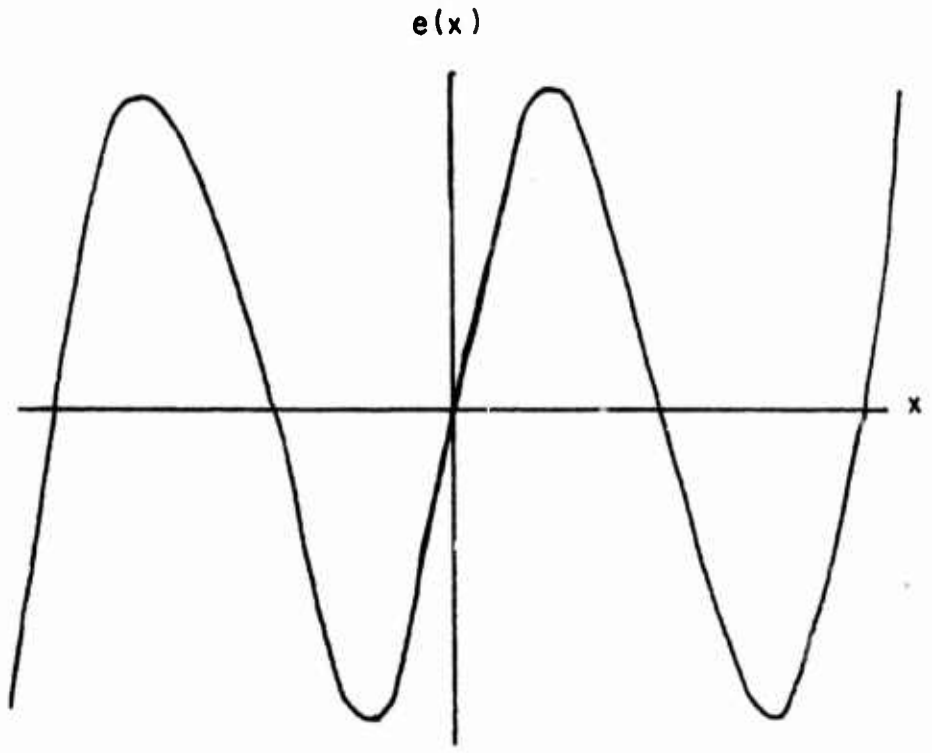


Figure 2. Error Function for Example 2

The algorithm was terminated when

$$|E(Q(k+1; x); X(k+1)) - E(Q(k; x); X(k))| \leq 0.0000009, \quad k = 0, 1, \ldots .$$

This took place at k = 1 for this example. The set X(0) was as follows:

$$X(0) = \{x_i : x_1 = -4, x_{i+1} = x_i + .4, i = 1, \ldots, 20\}.$$

The critical set for P(x) was (±4, ±2.895, ±0.92736).

Example 3

Function	$\ln(x+1)$
Range	$[0, 1]$
Approximation P(x)	$c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + c_5 x^5$
Number of Iterations	3
E(Q(2; x); X(2))	0.00000867
c_0	0.00000867
c_1	0.9993
c_2	-0.490747
c_3	0.286714
c_4	-0.133226
c_5	0.0311055

The initial sample points were (0.0, 0.08, 0.28, 0.5, 0.7, 0.9, 1.0). The critical set for P(x) was (0.0, 0.060052, 0.234368, 0.484917, 0.73004, 0.926248, 1.0). The algorithm was terminated when the difference between two successive errors was less than 0.0000001.

Example 4

Function	$f(x) = 0 \quad -1 \le x < 0$ $= 0.5 \quad x = 0$ $= 1 \quad 0 < x \le 1$
Range	$[-1, 1]$
Approximation P(x)	$c_0 + c_1x + c_2x^2 + c_3x^3 + c_4x^4 + c_5x^5$
Number of Iterations	3
E(Q(2; x); X(2))	0.5
c_0	0.5
c_1	4.25717
c_2	0.0
c_3	-12.641
c_4	0.0
c_5	9.38384

The algorithm was started with 21 equally spaced points in $[-1, 1]$, with the end points included. The search for the critical set required a step size of 0.0001. The critical set was (±1.0, ±0.8216, ±0.3673, ±0.00000001). In this case the maximum error took place at eight points rather than the expected seven. The proximity of two of the critical points makes the search for the critical set very difficult and extreme care is required.

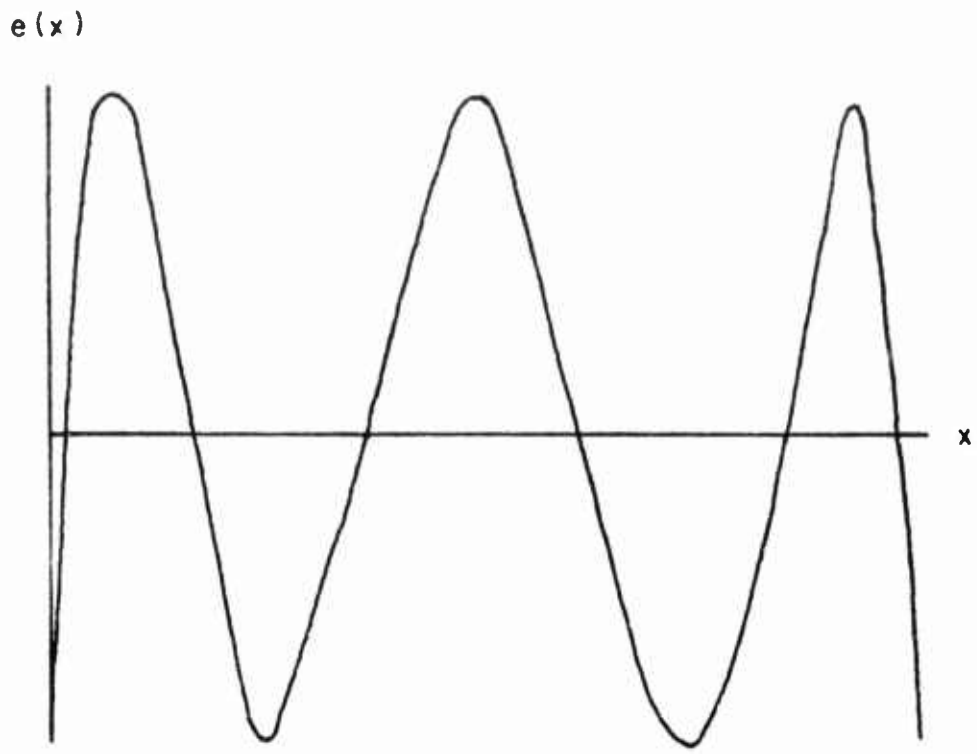


Figure 3. Error Function for Example 3

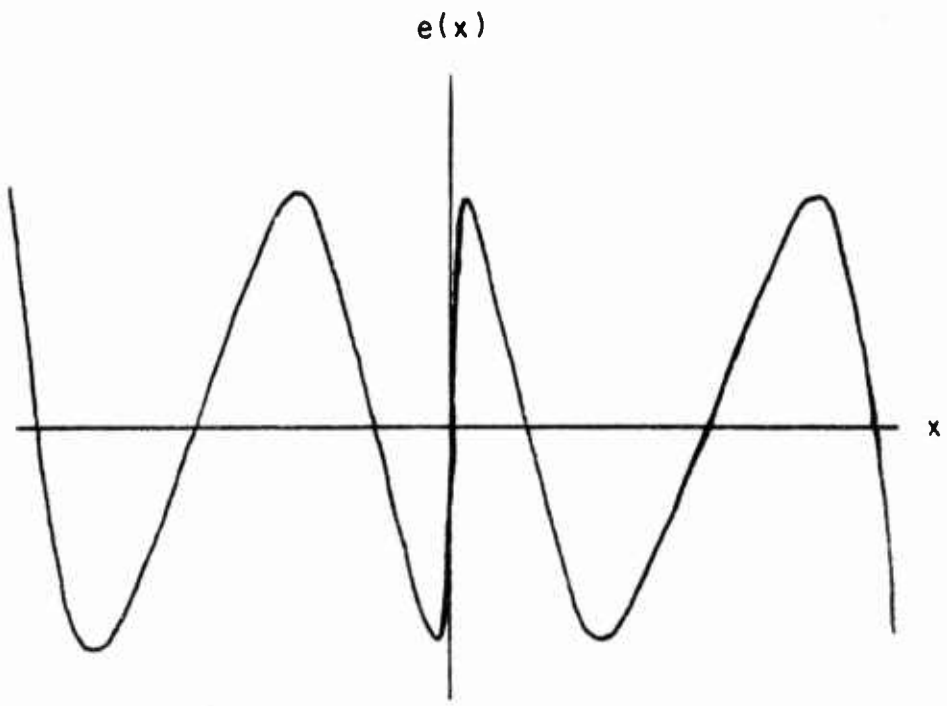


Figure 4. Error Function for Example 4

APPENDIX

COMPUTER PROGRAM

The program was coded entirely in the FORTRAN IV Programming Language. It consists of four main subroutines and five minor ones. Their functions are defined below:

RELAX - serves as an executive routine to the whole program and develops certain parameters necessary for the rest of the program.

APPROX - computes the polynomial of best approximation to the given function over a discrete point set.

SEARCH - determines the critical points and the direction in which to search for a new sample set. The search continues until three points straddle the required peak of the error curve. These three points and the corresponding values of the independent variable are used in PEAK.

PEAK - computes new critical point using the Lagrangian formula for a second degree interpolating polynomial. The three ordered pairs computed in SEARCH are used as data, and the derivative of the resulting polynomial is solved for its zero value to obtain the required point.

SIMPLX - used by APPROX to solve the linear program.

SETUP - used by APPROX to arrange the necessary input matrix to SIMPLX.

APPFN - used by APPROX to compute x_i^j, $i = 1, \ldots, m$
$j = 0, \ldots, n$

FF - computes values $f(x_i)$ of given function.

ERRFN - computes $f(x_i) - P(x_i)$, where P is the approximating polynomial.

Limitations on the program are (1) a maximum of 201 sample points and (2) a maximum degree of 9 on the approximating polynomials. These limitations are easily extended if desired.

The calling sequence to the program is as follows:

CALL RELAX (C, D, M, N, X, Y, ERRJ, MAXKT, DELTA, TOL, DTOL)

Output Parameters:

C = array of coefficients of desired polynomial

D = value of maximum error

ERRJ = error code
= 0, no error
= -1, approximation not found after allowable number of iterations
= -2 error did not increase on succeeding iteration

Input Parameters:

M = number of points in sample set

N = n + 1, where n is degree of polynomial

X = array of independent variable

Y = array of dependent variables $f(x_i)$

MAXKT = maximum number of iterations allowed

DELTA = step size to be used in search for succeeding critical sets

TOL = minimum allowable value of delta

DTOL = routine terminates when absolute difference between succeeding errors is less than DTOL

REFERENCES

1. Achieser, N. I. Theory of Approximation. New York: Frederick Ungar Publishing Co., 1965.

2. Blum, E., and Curtis, P. "Asymptotic Behavior of the Best Polynomial Approximation," Journal of the Association for Computing Machinery, Vol. 8, No. 4, 1961.

3. Cheney, E. Numerical Procedures for Chebycheff Approximations. MN264, Boeing Scientific Laboratory, 1962.

4. Cheney, E. Some Relationships Between the Chebycheff Approximations on an Interval and on a Discrete Subset of That Interval. MN262, Boeing Scientific Laboratory, 1962.

5. Dantzig, G. B. Linear Programming and Extensions. Santa Monica: The Rand Corporation, 1963.

6. Frank, W. L., and Curtis, P. C. "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.

7. Gass, S. Linear Programming, Methods and Applications. New York: McGraw-Hill, 1958.

8. Hamming, R. W. Numerical Methods for Scientists and Engineers. New York: McGraw-Hill, 1962.

9. Hastings, C., Jr. Approximations in Numerical Analysis. Santa Monica: The Rand Corporation, 1953.

10. Hastings, C., Jr. The Incomplete Approximator. Santa Monica: The Rand Corporation, 1953.

11. Jackson, Dunham. The Theory of Approximation. New York: The American Mathematical Society, 1930.

12. Kelley, J. "The Application of Linear Programming to Curve Fitting," Journal of the Society for Industrial and Applied Mathematics, Vol. 6, No. 1, 1958.

13. Korovkin, P. P. Linear Operators and Approximation Theory. Delhi: Hindustan Publishing Co., 1959.

14. Lambert, P. V., and Neuts, M. F. Application de la Theorie des Jeux aux Problemes d' Approximation. Rando Beach: Space Technology Laboratories, 1963.

15. Langer, R. E. (ed.). On Numerical Approximation. Madison: University of Wisconsin Press, 1959.

16. Milne, W. E. Numerical Calculus. Princeton: Princeton University Press, 1949.

17. Motzkin, T. S., and Walsh, J. W. Polynomials of Best Approximation on a Real Finite Point Set. Los Angeles: Numerical Analysis Research, University of California, 1957.

18. Murnaghan, F. D., and Wrench, J. W. The Approximation of Differentiable Functions by Polynomials. David Taylor Model Basin, Research and Development Report 1175, 1958.

19. Poussin, C. de la Vallee, Lecons sur l Approximations des Fonctions d'une Variable Reels. Paris: Gauthier Villars, 1919.

20. Shenitzer, A. "Chebycheff Approximation of a Continuous Function by a Class of Function," Journal of the Association for Computing Machinery, Vol. 4, 1957.

21. Speilberg, K. "Representation of Power Series in Terms of Polynomials, Rational Approximations, and Continued Fractions," Journal of the Association for Computing Machinery, Vol. 8, No. 4, 1961.

22. Stiefel, E. L. An Introduction to Numerical Mathematics. New York: Academic Press, 1963.

23. Todd, John (ed.). A Survey of Numerical Analysis. New York: McGraw-Hill, 1962.

24. Walsh, J. L. Interpolation and Approximation. New York: The American Mathematical Society, 1935.

25. Ward, L. E. "Linear Programming and Approximation Problems," American Mathematical Monthly, Vol. 68, No. 1, 1961.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Systems Engineering Group
Wright-Patterson Air Force Base, Ohio

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b GROUP

3. REPORT TITLE

A METHOD FOR CALCULATING POLYNOMIALS OF BEST APPROXIMATION

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(Last name, first name, initial)*

Oliver, Alfred

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
August 1966	29	25

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	SEG-TR-66-32
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Systems Engineering Group Wright-Patterson Air Force Base, Ohio

13. ABSTRACT

This report deals with minimax approximations to functions defined on the real line. These approximations are of particular use to mathematicians having access to a digital computer facility.

The cases considered are for functions defined as discrete points on the Cartesian plane and for continuous, bounded functions defined on a closed interval of the real line. A method for obtaining the approximations using linear programming techniques is presented for each case.

An analysis of the error function and a brief outline of the computer program are included.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Mathematics Numerical Analysis Approximation Theory Chebycheff Approximation Linear Programming						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through

_____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through

_____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

_____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.