UNCLASSIFIED

AD NUMBER
AD248573

LIMITATION CHANGES

TO:
Approved for public release; distribution is unlimited.

FROM:
Distribution authorized to U.S. Gov't. agencies and their contractors;
Administrative/Operational Use; SEP 1960. Other requests shall be referred to Air Force Cambridge Research Laboratories, Hanscom AFB, MA.

AUTHORITY
AFCRL ltr, 3 Nov 1971

THIS PAGE IS UNCLASSIFIED
NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. 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.
Best Available Copy
AN ALGORITHM FOR FINDING RATIONAL APPROXIMATIONS

H. F. Mattson, Jr.

SEPTEMBER 1960

ELECTRONICS RESEARCH DIRECTORATE
AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
AIR FORCE RESEARCH DIVISION (ARDC)
UNITED STATES AIR FORCE
BEDFORD MASSACHUSETTS
AN ALGORITHM FOR FINDING RATIONAL APPROXIMATIONS

H. F. MATTSON, Jr.

Project 5632
Task 56321

SEPTEMBER 1960

COMPUTER AND MATHEMATICAL SCIENCES LABORATORY
ELECTRONICS RESEARCH DIRECTORATE
AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
AIR FORCE RESEARCH DIVISION(ARDC)
UNITED STATES AIR FORCE
BEDFORD MASSACHUSETTS
NOTE

As I was preparing the manuscript of this Report for typing, I discovered, on April 22, 1960, that my method of approximation, at least when restricted to polynomials, is not essentially new. A very similar method is described in [6]. For a comparison of the two methods, see §5.
§1. ABSTRACT

Scientific work frequently requires numerical values of functions. Although much general information is often known about these functions, values of them are nevertheless often difficult to compute; although in almost all cases some method for making this computation, however lengthy it may be, is known. The purpose of a rational approximation to a function is to provide a rapid and convenient way to calculate numerical values of the function to within a predetermined error. This paper considers the question of how to find rational approximations to given functions. In §2 there appear definitions of terms, a precise statement of what our criterion of best fit is, and statements of some classical results. In §3, two closely related iterative methods for finding best rational approximations are defined. In §4, a proof of convergence of these methods is given for a special case (in which both methods are the same). In §5, these methods are compared with some others. In §6, some results obtained by one of the methods of §3 are presented, together with a brief description of the computer program used to obtain them.

§2. DEFINITIONS AND KNOWN RESULTS

A few preliminary definitions are necessary to this discussion. We shall restrict ourselves to the finite interval \( I = [a, b] \) on the real line. If we consider the space of all continuous functions (with real values) defined on \( I \), it is natural and commonplace...
to define the distance between two such functions $h$ and $g$ as the maximum absolute value of the difference $h - g$:

$$\max_{a \leq x \leq b} |h(x) - g(x)| = \|h - g\|.$$ 

In general, we define $\|G\|$, for any continuous function $G$ defined on $I$, as

$$\|G\| = \max_{a \leq x \leq b} |G(x)|.$$ 

Throughout this paper $f$ will denote a fixed, continuous function over $I$. (We shall impose an additional restriction on $f$ at one point later on.) For given non-negative integers $m$ and $n$, we consider the family $F \cap G$ of all rational functions of the form $S(x) = P(x)/Q(x)$, where $P(x)$ is a polynomial of degree at most $m$, and $Q(x)$ is a polynomial of degree at most $n$ having no zeros in $I$. Each $S$ in $F \cap G$ is at a certain distance $d_S = \|S - f\| \geq 0$ from $f$. The set of numbers $d_S$, with $S$ in $F \cap G$, has a greatest lower bound $d$, a notation which will be fixed throughout this paper. The questions on rational approximation which naturally arise are the following: Is there an $R$ in $F \cap G$ such that $d_R = d$? If so, what more can we say about $R$? In particular, how can we find it?

The answer to the first question is yes: There exists an $R$ in $F \cap G$ such that $d_R \leq d_S$ for every $S$ in $F \cap G$ \cite[1, p. 53]{1}. For this $R$, then, $d_R = d$. Furthermore, such $R$ is unique; that is, if $S$ is any rational function in $F \cap G$ different from $R$, then $d_R < d_S$ \cite[1, p. 56]{1}. $R$ will always denote this best approximation. "Best", or "best-fitting", is here used in the sense previously defined; it is often called "best in the sense of Chebyshev" in the literature.

We now quote two important theorems on rational approximations which will give us more information about $R$ and $d$. (No pun intended) The first theorem will allow us to find lower bounds for $d$; the second characterizes $R$ in terms of some properties which will prove to be useful later on.

\[1\) Numbers in square brackets refer to the bibliography at the end of this Report.
Let $R_0$ be any rational function in $F_{m,n}$ so $R_0(x) = P_0(x)/Q_0(x)$, with

$$P_0(x) = a_\mu x^{m-\mu} + \ldots + a_m, \quad Q_0(x) = b_\nu x^{n-\nu} + \ldots + b_n, \quad b_\nu \neq 0, \quad 0 \leq \mu \leq m, \quad \text{and} \quad 0 \leq \nu \leq n.$$

Define $N = m + n + 2 - \delta$, where $\delta = \min(\mu, \nu)$.

Assume also that $R_0(x) \neq 0$ and that $P_0(x)$ and $Q_0(x)$ have no common divisor.

**THEOREM A.** Suppose that the error function $E_0 = R_0 - f$ at some $N$ points $x_1 < \ldots < x_N$ in $I$ assumes respectively the values $-v_1, +v_2, -v_3, \ldots, (-1)^N v_N$ different from zero and of alternating signs (thus all $v$'s have the same sign).

If $R_1$ is any member of $F_{m,n}$ with error function $E_1 = R_1 - f$, then

$$d_{R_1} = \| E_1 \| \geq \min \{|v_1|, \ldots, |v_N|\}.$$

(If $P_0 = 0$, then the same inequality holds with $N = n + 2$.)

A consequence of this theorem is that $d = d_{R_1} \geq \min \{|v_1|, \ldots, |v_N|\}$.

And if $R_1$ is any member of $F_{m,n}$, the error function of which has sufficiently many extreme values of alternating signs, then $d$ is not less than the minimum of these extreme values (in magnitude).

**THEOREM B.** $R_0$ is the best-fitting rational function $R$ if and only if there exists at least $N$ points $x_1 < \ldots < x_N$ in $I$ at which $E_0 = R_0 - f$ assumes the values $E_0(x_i) = (-1)^{i+\alpha} \| E_0 \|, \quad i = 1, \ldots, N$, where $\alpha$ is either 0 for all $i$ or 1 for all $i$.

($R = 0$ is the best-fitting rational function in $F_{m,n}$ if, and only if, there are at least $N = n + 2$ points $x_1 < \ldots < x_N$ for which $f(x_i) = (-1)^{i+\alpha} \| f \|$.)

For the proof of Theorem A, see [1, pp. 52-53]. A proof of Theorem B also occurs in [1, pp. 55-57].

In proving the uniqueness of $R$ [1, pp. 56-57], mentioned earlier, one uses Theorem B.
We shall use the terms "extremum" and "extreme value" as follows: An extremum of the function \( g \) is a point in \( I \) at which \( g \) takes an extreme value. If \( R_0 \notin F_{m,n} \), we define an admissible set of extrema of \( E_0 \) (\( E_0 = R_0 - f \)) to be \( N \) points \( x_1 < \ldots < x_N \) in \( I \) such that:

1) each \( x_i \) is an extremum of \( E_0 \), and

2) \( E_0(x_1) \neq 0 \) and alternates in sign as \( i \) increases from 1 to \( N \).

The set \( \{ E_0(x_i) ; i = 1, \ldots, N \} \) is then called an admissible set of extreme values of \( E_0 \).

Finally, let me observe that the only extreme (of any function \( G \)) pertinent to the situation under discussion in this Report are the maxima of \( |G| \). Therefore "extrema" should be tacitly so understood here.

§ 3. DEFINITIONS OF ALGORITHMS

We now define two iterative procedures which in some cases (see § 6) are known to converge to the best-fitting rational approximation to our given continuous function \( f \) over \( I = [a, b] \). (There are no cases known to me in which the procedures do not converge to \( f \), but a proof of convergence is known to me only for the restriction to \( n = 0, f \in C^0 \).)

We shall call our first algorithm the "non-linear" algorithm and our second the "linear" algorithm. Both have in common the following first step:

1.° For a given \( m, n \) select (say, by interpolation at the estimated zeros of \( E \)) \( R_0 \notin F_{m,n} \) such that \( E_0 = R_0 - f \) has an admissible set of extrema.

Nonlinear Method:

2.° Given \( R_{j-1} \notin F_{m,n} \), such that \( E_{j-1} = R_{j-1} - f \) has an admissible set of extreme \( x_1, \ldots, x_N \), determine \( R_j \in F_{m,n} \) by imposing the \( N \) conditions...
\[ R_j(x_i) - f(x_i) = (-1)^i y_j \quad i = 1, \ldots, N, \] where \( y_j \) is an unknown. The other unknowns are, of course, the coefficients occurring in \( P_j \) and \( Q_j \) in \( R_j \). 2)

**Linear Method:**

2°. Given \( R_{j-1} = P_{j-1}/Q_{j-1} \in F_m, n \geq 0 \), such that \( E_{j-1} = R_{j-1} - f \) has an admissible set of extrema \( x_1 < \ldots < x_N \), determine \( R_j = P_j/Q_j \in F_m, n \) by the \( N \) conditions

\[ P_j(x_i) - f(x_i) Q_j(x_i) = (-1)^i Q_{j-1}(x_i) y_j, \quad i = 1, \ldots, N. \]

Choose the leading coefficient of each \( Q_j \) to be 1. Notice that these equations are linear in the unknowns \( y_j \) and the coefficients occurring in \( P_j \) and \( Q_j \). 2)

**Comments on these methods:**

The motivation for the non-linear method clearly comes from the characterization of the best-fitting rational function in Theorem B. The desired best-fitting rational function is obviously a fixed point of the transformation defined in Step 2°. (modulo the complications mentioned in footnote 2). The linear method is derived from the non-linear one in an obvious way, and it also has the desired best-fitting rational function as a fixed point (modulo the complications mentioned in footnote 2).

It is not known whether either of those methods actually stays inside \( F_{m,n} \) in general. That is, for some \( j \), \( R_j \) as defined in either method may, \textit{a priori}, fail to exist. Also, for some \( j \), \( R_j \) may, \textit{a priori}, have a pole in \( I \). If one satisfied these pre-conditions, the central question would remain, namely, whether one or the other method converges. These three questions appear to be difficult for the general case.

2) It may happen that \( R_j - f \) has more than \( N \) extrema. In such a case I require the choice of a particular admissible set of extrema. I have stated the method in its simplest form here, for clarity; it appears in full generality as step 2° in §4.
(i.e., m and n arbitrary, f any continuous function on I) of the non-linear method and more difficult for that of the linear method. But in the case n = 0 both methods coincide, and a considerable number of the necessary results can be proved for arbitrary f. Finally, if one then restricts f to have m continuous derivatives, a full existence and convergence proof is possible. We present this proof in the next section.

§ 4. PROOFS FOR THE CASE n = 0

We now confine ourselves to polynomial approximations $P_0$ to our given f. Both methods are the same in this case. Our first step is to satisfy the pre-conditions by proving the existence of $P_{j+1}$, given a $P_j$ having the required properties. The proof leads naturally to further results, all of which we include in the following theorem.

THEOREM 1. Let $P_0$ be a polynomial, either 0 or degree $m - \mu$, where $0 \leq \mu \leq m$, such that there are $N = m + 2$ points

$$(a \leq ) x_1 < \ldots < x_N (\leq b)$$

at which the error function $E_0 = P_0 - f$ takes on respectively the non-zero values

$E_0(x_i) = (-1)^i v_i$, $i = 1, \ldots, N$, where all $v_i$ have the same sign. Then there exists a polynomial $P_1$ of degree at most m, and a number y such that the error function $E_1 = P_1 - f$ takes the values

$E_1(x_i) = (-1)^i y$ \text{(4.1)},

and

a) y is uniquely determined by the condition (4.1),

b) If not all $v_i$ are equal, then

$$\min_{1 \leq i \leq N} |v_i| < |y| < \max_{1 \leq i \leq N} |v_i| \quad (\text{therefore } y \neq 0), \quad \text{(4.1)}$$

c) Sign y = sign $\{v_i\}$. 

-6-
PROOF. We are given \( N \) distinct points \( x_i \) at which there hold the equations

\[
P_0(x_i) - [f(x_i) + (-1)^i v_i] = 0, \quad i = 1, \ldots, N.
\]

These equations may be thought of as \( N \) homogeneous equations, already solved, for the \( N-1 \) coefficients of an \( m \)-th degree polynomial, plus one more "unknown". That is, \((0, \ldots, 0, a_\mu, a_{1+\mu}, \ldots, a_m, 1)\) is the solution-vector (transposed), and the \( N \times N \) coefficient-matrix is

\[
M = \begin{bmatrix}
x_1^m & x_1^{m-1} & \cdots & 1 & -[f(x_1) + v_1] \\
x_2^m & x_2^{m-1} & \cdots & 1 & -[f(x_2) + v_2] \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
x_N^m & x_N^{m-1} & \cdots & 1 & -[f(x_N) + v_N]
\end{bmatrix}.
\]

The existence of \( P_0 \) with the given properties imply that the determinant of the matrix \( M \) is zero.

What we first wish to prove is that there is a number \( y \) such that the matrix we obtain by replacing each \( v_i \) by \( y \) in \( M \) is also zero. (Such a \( y \) would imply the existence of the desired \( P_1 \).) To this end we expand \( \det(M) \) by cofactors of the last column, obtaining

\[
\sum_{i=1}^{N} (-1)^{N+1} \sigma_i [f(x_i) + (-1)^i v_i] = 0 \quad (4.2);
\]
where the minor $e_i$ is the van der Monde determinant

$$
e_i = \begin{vmatrix} x_1^m & x_2^{m-1} & \cdots & 1 \\
\vdots & \vdots & & \vdots \\
x_1^m & x_1^{m-1} & \cdots & 1 \\
\vdots & \vdots & & \vdots \\
x_N^m & x_N^{m-1} & \cdots & 1
\end{vmatrix}$$

(this row omitted);

thus we have, by the well known formula,

$$e_i = \prod_{j < k} (x_j - x_k).$$

Since the $x_1$'s are all distinct, each $e_i \neq 0$; since the $x_1$'s are arranged in increasing order, all $e_i$ have the same sign, each one being the product of the same number of negative factors. Having noted these properties of the $e_i$'s, we now rewrite equation (4.2) as

$$e_1 v_1 + \ldots + e_N v_N = - \sum_{i=1}^{N} (-1)^i e_i f(x_i) = df e.$$  \hspace{1cm} (4.3)

This shows that $|e| = \sum |e_i v_i| > 0$ and that sign $e = (\text{sign } e_1) (\text{sign } v_i)$. We can now immediately satisfy our requirements for the existence of $P_1$ by choosing $y$ to satisfy

$$e_1 y + \ldots + e_N y = e,$$  \hspace{1cm} (4.4)

of $y \in \sum_{i=1}^{N} e_i$. Furthermore, this is the only choice open to us for $y$. 

-8-
A comparison of (4.2) and (4.3) shows that \( \min |v_1| < |y| < \max |v_1| \) unless all \( v_1 \) are equal (in which case \( y = v_1 \)); and, finally, \( \text{sign } y = \text{sign } e \cdot \text{sign } e_1 = \text{sign } v_1 \). QED.

There are two points to notice about this theorem. One is that we do not require the \( x_1 \)’s to be extrema of \( E_0 \) but only to be points where \( E_0 \) alternates in sign. In the application of this theorem to our iterative process, however, we shall take them as extrema.

The other point is that although \( P_0 \) may have degree less than \( m \), the same is not necessarily true of \( P_1 \), as the following example shows:

Let \( f(x) = \cos \pi x \), over \( I = [a, b] = [0, 5/2] \). Take \( m = 2 \), so that \( N = 4 \). For \( P_0 \), take \( P_0(x) = a_0 \), with \( 0 < a_0 < 1 \). Then we have

\[
\begin{align*}
  x_1 &= 0 & x_2 &= 1 & x_3 &= 2 & x_4 &= 5/2 \\
  v_1 &= 1-a_0 & v_2 &= 1+a_0 & v_3 &= 1-a_0 & v_4 &= a_0 ,
\end{align*}
\]

as is obvious from the following sketch:

This same sketch makes it obvious that there is no straight line \( u = P_1(x) = a_1 x + a_1 \) having deviations of equal magnitude at the \( x_1 \)’s. Therefore \( P_1 \) will be a parabola, something like the one sketched with a dashed line.
What we have proved in Theorem 1 is that we can always construct our sequence \( \{P_j\} \) satisfying the conditions of step 2°, provided there exists a first polynomial \( P_0 \) satisfying the condition of step 1°. But we can always find \( P_0 \) by solving \( P_0(x) - f(x) = (-1)^n y_0 \) for the unknown coefficients in \( P_0 \) and the unknown \( y_0 \), for any distinct \( x_1, \ldots, x_N \in I \). The existence of this \( P_0 \) and \( y_0 \) is given by the proof of Theorem 1, in which \( y_0 \) is given by (4.4) with \( e \) defined by (4.3). It is only necessary to choose the \( x_i \) so that \( e \neq 0 \), the possibility of which follows, in case \( f \) is not a polynomial of degree at most \( m \), from the existence of the best-fitting polynomial and the consequent existence of \( v_i \)'s with alternating signs.

Having shown the existence of our sequence \( \{P_j\} \), we now show that, under suitable restrictions, it converges to \( P \).

We first must modify our definition of step 2° to take account of the possibility that there are more than one set of admissible extrema. We prove a simple lemma:

**Lemma 1.** If \( E_j = P_j - f \) has an admissible set of extrema, then there is an admissible set of extrema values containing \( \| P_j - f \| = \#d_j \).

**Proof.** Let \( \#d_j \) occur as a value of \( E_j \) at \( x \). If \( x \) is already in the admissible set of extrema, then we are done. If \( x \) lies between two of the extrema \( x' \) and \( x'' \), then \( E_j(x) \) must have the same sign as one of \( E_j(x') \) and \( E_j(x'') \). Replace that one by \( E_j(x) \). If \( x \) lies entirely to the left of the admissible extrema, then either replace the one nearest \( x \) by \( x \), or delete the farthest one and include \( x \), all depending on whether or not the sign of \( E_j \) at the one nearest \( x \) is the same as that of \( E_j(x) \). QED.

---

3) I am indebted to Novodvorskii and Pinsker, [4], via Shenitzer [6], for this point. It is slightly easier to prove the possibility of this than to show the possibility of avoiding a tangency of \( P_0 \) and \( f \) when interpolating to \( f \) at the estimated zeros of \( E \), which I had suggested earlier. In order to obtain linearity of the equations for \( R_0 \), when \( n > 0 \), however, one needs to interpolate as first suggested.
It follows immediately from this lemma that if we choose a set of admissible extreme values having the largest possible minimum (in magnitude), then we may replace it by a set containing \( \| E_j \| \) without changing the magnitude of the minimum extreme value in it. It is this procedure that we follow when there is more than one choice of a set of admissible extreme values at any stage. Accordingly, we substitute the following step 2'° for step 2°:

2°. Given \( P_j \in F_{m,0} \) such that \( P_j - f \) has a set of \( (N) \) admissible extreme values, choose a set \( S \) of admissible extreme values containing the largest possible minimum (in magnitude) and containing \( \| P_j - f \| \). Let \( x_1, \ldots, x_N \) be the extrema corresponding to this set \( S \) (i.e., \( S = \{ E_j(x_i) : i = 1, \ldots, N \} \)). Determine \( P_{j+1}(x) = a_0 x^m + \ldots + a_m \) by imposing the \( N \) conditions

\[
P_{j+1}(x_i) - f(x_i) = (-1)^i y_{j+1}, \quad i = 1, \ldots, N,
\]

where \( y_{j+1} \) is an unknown.

The possibility of carrying out step 2° has been proved in Theorem 1 and Lemma 1.

We now turn to the proof of convergence.

**Lemma 2.** The sequence \( \{ | y_j | \} \) determined by step 2° is strictly increasing and bounded above by \( d = \| E - f \| \); (unless some \( P_j = P \); then all \( | y_k | = d \) for \( k \geq j \)).

**Proof.** We are given that \( | y_j | = E_j(x_i) \), where the \( x_i \) are the extrema belonging to the set \( S \) of extreme values of \( E_{j-1} \) defined in step 2°. Let \( V_1 = E_{j-1}(x_i) \). Let \( S' = \{ (-1)^{x_i+1} v_i \} \) be a set of extreme values of \( E_j \) as prescribed by step 2°. Since there is an admissible set of extreme values of \( E_j \) such that \( | y_j | \) is smaller (in magnitude) than all of them, it follows that

\[
| y_j | \leq \min | v_i |.
\]  

(4.5)
Now (4.5) plus Theorem 1 yields \( |y_j| \leq \min |v_i| < |y_{j+1}| \). And finally, Theorem A gives \( |y_j| \leq \min |v_i| \leq \|E\| = d \). QED.

We now prove a convergence theorem.

**THEOREM 2.** Every convergent subsequence of \( \{P_j\} \) has limit \( P \).

**PROOF.** Every subsequence \( \{P_{j_k}\} \), convergent or not, has the property that the sequence of attached \( |y_{j_k}| \) converges to some limit \( d' \leq d \). Let \( P^* \) be the limit \( \lim_{k \to \infty} P_{j_k} \) of our convergent subsequence. Then \( P^* \) is a polynomial of degree at most \( m \), (because the sequences of coefficients of the \( P_{j_k} \) are bounded; therefore there is a subsequence of the subsequence \( P_{j_k} \) which has a polynomial \( p(x) \) as limit. But since the subsequence \( P_{j_k} \) is already convergent, the limit \( P^* \) must be the polynomial \( p(x) \) just mentioned. Finally, the coefficients of the \( P_{j_k} \) are bounded because the polynomials are bounded at \( m + 1 \), in fact at all, points of 1.) Thus \( P^* \) is the uniform limit of \( \{P_{j_k}\} \).

We shall prove that \( d' = \|P^* - f\| \). First, \( EI = P^* - f \) has an admissible set \( S \) of extreme values, since \( P^* \) is the uniform limit of \( \{P_{j_k}\} \); and, for the same reason, the set \( S \) has \( d' \) as minimum magnitude, in view of the inequalities (4.1) of Theorem 1. Therefore we may, and do, apply step 2'0 to \( P^* \), obtaining \( P^{**} \). If \( \|E^*\| > d' \), then the value of the \( y^* \) found in step 2'0 (satisfying \( E^{**} \) (extrema of \( E^* \) = \( \pm y^* \)) would satisfy \( y^* > d' \). We choose \( k \) large enough so that corresponding admissible extrema of \( E_{j_k} \) and \( E^* \) are close enough to each other to yield \( |y_{j_k}| \) so close to \( |y^*| \) that \( |y_{j_k}| > d' \). This is possible since \( y \), determined by formula (4.4), depends continuously on the extrema. But this result contradicts the definition of \( d' \). Therefore \( \|E^*\| = d' \leq d \). But since we always have \( \|E^*\| \geq d \), we have proved that \( d = d' \).

By the uniqueness of \( P \), we conclude that \( P^* = P \). QED.
Now we are almost finished, for if we knew that \( \{P_j\} \) were a bounded sequence, we could conclude from the previous theorem that \( \{P_j\} \) converged to \( P \). 4) We now introduce a hypothesis, probably stronger than necessary, which implies the boundedness of \( \{P_j\} \).

**THEOREM 3.** If \( f \) has \( m \) continuous derivatives, then the sequence \( \{P_j\} \) constructed according to either step 2° or step 2'° is bounded over \( I \).

**PROOF.** Each error function \( E_j = P_j - f \) has at least \( N - 2 = m \) distinct extrema interior to \( I \). The existence of a continuous first derivative implies that \( E_j' \) has at least \( m \) distinct zeros inside \( I \). Therefore \( E_j' \) has at least \( m - 1 \) distinct extrema and \( E_j'' \) an equal number of zeros inside \( I \). Continuing in this way, we find that \( E_j^{(m)} \) has at least one zero inside \( I \). Therefore we have \( P_j^{(m)}(z') = m! a_{oj} = f^{(m)}(z') \) for some \( z' \) in \( I \), where \( a_{oj} \) is the leading coefficient of \( P_j \). Since \( f^{(m)} \) is assumed continuous on \( I \), it is founded there, from which it follows that \( \{P_j^{(m)}\} \) is bounded.

From the relations
\[
P_j^{(\alpha - 1)}(x) = \int_{z}^{x} P_j^{(\alpha - 1)}(t) \, dt + f^{(\alpha - 1)}(z), \quad \alpha = m, m - 1, \ldots, 1,
\]
where \( z \) is a zero of \( E^{(\alpha - 1)} \), we conclude inductively that \( \{P_j\} \) is bounded, since all derivatives of \( f \) are bounded, for order not greater than \( m \).

Incidentally, we could conclude easily from the above proof that the coefficients of the \( P_j \) are all bounded; \( \{a_{oj}\} \) is bounded since \( P_j^{(m)} = m! a_{oj} \) is bounded; one integration introduces the \( a_{lj} \) which are therefore bounded, and so on. But we already know the boundedness of the coefficients follows from that of the polynomials, in general.

---

4) We would have: In a complete metric space (here the Euclidean \((m + 1)\)-space of coefficients) a bounded sequence with at most one limit point converges.
We summarize the import of this section in the following theorem.

**THEOREM 5.** Let \( f \in C^m \) on \( I \), and let the sequence \( \{ P_j \} \) of polynomials be defined by the rules: \( P_0 = f \) has an admissible set of extrema; each \( P_j, \ j > 0 \), is obtained from \( P_{j-1} \) by the rule of step 2°. Then \( \{ P_j \} \) converges uniformly on \( I \) to \( P \), the polynomial of degree at most \( m \) which lies nearest to \( f \) in the matrix defined in § 2.

§ 5. COMPARISONS WITH OTHER METHODS

The methods of Remez [5], represented in some fashion in [4], is described in English in 6 for approximations of the form \( s(x) P(x) \), where \( s \) is a fixed continuous function with no zeros in \( I \), and \( P \) is a polynomial. Proofs, said to be given in [4], are omitted from [6]. The latter reference is the only one I have been able to read to date.

We now describe Remez's method in our terminology:

The method begins by the choice of the initial approximation \( P_0 \) as described in § 4: For \( x_1 < \ldots < x_N \in I \), we determine \( P_0 \) by the \( N \) equations \( P_0(x_i) - f(x_i) = (-1)^i y_0 \). Step 2. Let \( x' \) be a point of \( I \) at which \( E_0 = P_0 - f \) takes the value \( \pm \| E_0 \| \). Replace one of the \( x_i \) by \( x' \), calling the resulting points \( x_{11} < x_{21} < \ldots < x_{N1} \) in such a way that they are an admissible set of extrema of \( E_0 \) (Cf. Lemma 1.) Now determine \( P_1 \) by solving the \( N \) equations \( P_1(x_{11}) - f(x_{11}) = (-1)^i y_{11}, \ i = 1, \ldots, N \). Find \( P_2 \) by replacing one of the \( x_{11} \) by \( x' \) such that \( E_1(x') = \pm \| E_1 \| \), and so on.

It is obvious that the conclusions of Theorem 1 hold here also, and that the \( y_j \)'s of Remez increase monotonically (in magnitude) to \( d \). The rest of our proof of convergence clearly carries over to this process without essential change. The presence of the function \( s \) would complicate the proofs in no essential way; \( s \) was omitted from the present report chiefly for reasons of clarity.
As a practical matter, the present method should converge faster than that of Remez, since the $|\gamma_j|$ are larger than the corresponding quantities in Remez's method. This quicker convergence is paid for by a more complicated choice of the admissible extrema $x_j$ at each stage, however, except for the case when there are always exactly $N$ admissible extrema. In this important special case, there is no more difficulty in carrying out the present method than there is in doing that for Remez's; for in order to find $||E||$ one must find all extreme values of $E_j$.

If the computation of $f(x)$ were difficult, then the method of the present report might well be preferable to that of Remez.

I was led to the method of this report partly by ruminating on Hastings's method, as defined by P. W. Ketchum in "Mathematical Reviews" (Hastings's book suffers from a certain lack of definition). For me, the central point of both Hastings's method and my own is the "iterative assumption" of stability of the extrema of $E_j$. This point plus Theorem B led me to my method. I then noticed the following comparison: Hastings linearizes the $N$ non-linear equations in step $2^\circ$ by assigning a numerical value to $y_j$, thus reducing the number of unknowns to $N-1$; he solves $N-1$ equations and hopes for a correct value at the $N$-th extrema. I linearize these equations by replacing the unknown coefficient of $y_j$ by a known number, thus preserving the number of unknowns.

§6. EXPERIMENTAL DATA

The so-called "linear algorithm" of §4 for finding rational approximations has been programmed, with $m = n = 2$, for the Cambridge Computer. The following are the functions $f$ approximated, the interval $I$, the value of $||E||$ obtained, the smallest maximum value of $|E_k|$ and the number $k$ of times Step $2^\circ$ was performed in order to obtain the final approximation:

5) I wish to acknowledge most gratefully the kind assistance of Miss Helen Willett in obtaining these results from the Cambridge Computer.
In each of the above three "experiments", there were exactly \( 6 = N \) extrema of \( E_j \) for each \( j \); the extreme values always alternated properly in sign. These extrema were maxima of \( |E_j| \). Two extrema were always at the end-points of \( I \).

The minimum magnitude of the extreme values is shown in order to provide an idea of the closeness of the last approximation \( R_k \) to the best approximation \( R \). In the cases \( \exp \) and \( \log \), these numbers and those under the heading \( \|E\| \) are probably not accurate to the three significant figures presented, since they are the last three significant figures of the eight available on the Cambridge Computer.

There follow the values of the coefficients of the approximations \( R_k \) discussed above. In each case \( R_k \) has the form

\[
R_k(x) = \frac{a_2 x^2 + a_1 x + a_0}{x^2 + b_1 x + b_0}.
\]

<table>
<thead>
<tr>
<th></th>
<th>( a_2 )</th>
<th>( a_1 )</th>
<th>( a_0 )</th>
<th>( b_1 )</th>
<th>( b_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \exp )</td>
<td>1.1045366</td>
<td>6.5455741</td>
<td>12.869739</td>
<td>-6.3197327</td>
<td>12.868806</td>
</tr>
<tr>
<td>( \log )</td>
<td>3.3615004</td>
<td>1.7750524</td>
<td>-5.3365379</td>
<td>5.6992101</td>
<td>1.9999023</td>
</tr>
<tr>
<td>( \text{sine} )</td>
<td>.88630520</td>
<td>-8.4079627</td>
<td>18.363517</td>
<td>-9.0106829</td>
<td>21.528251</td>
</tr>
</tbody>
</table>

Some tests of significance of digits in the coefficients of the approximation to \( \exp \) were made. It was found that rounding the coefficients to seven significant figures produced a spread of \( 0.25 \times 10^{-5} \) between minimum and maximum extreme values, whereas rounding to six decimal places produced a spread of \( 0.03 \times 10^{-5} \), the maxi-
mum being $8.701 \times 10^{-5}$, the minimum $8.67 \times 10^{-5}$. These numbers of course are potentially in error because of round-off in the machine, but they indicate that the last-mentioned rounding produces a result just as good as the original.

A brief description of the program follows. Letting $h$ denote $(b - a)/20$, the computer finds $R_0$ by interpolating to $f$ at the five equally spaced points $x_1 < \ldots < x_5$, where $x_1 = a + h$ and $x_5 = b - h$. The extrema of $E_0$ are then found; at every stage the program assumes that $a$ and $b$ are extrema of $E_j$; the interior extrema are found by solving $E_j'(x) = 0$ by the method of regula falsi. The equations of step 2 are then solved; and the process is repeated. The criterion for stopping at $j = k$ is that the extrema of $E_k$ be not too different from the corresponding extrema of $E_{k-1}$, which are stored at each stage. The coefficients $a_k$ and $b_k$ of each $R_j$ are printed, and at the end the values of the extrema of $E_k$ and the corresponding extreme values are printed.

The arbitrariness of the above procedure for finding $R_0$ can lead to difficulty. In particular, for $f = \sin$, it gave an $R_0$ having poles in $I$ for $I = [0.1, 6.8]$ and $I = [0.5, 6.9]$. Up to now, the method has "converged"; however, so long as $R_0$ had no poles in $I$. As now programmed, however, it would probably fail if some $E_j$ had more than $N = 6$ extrema, or if $a$ (or $b$) were a minimum of $|E_j|$, rather than a maximum.

At no point, except in the non-essential final print-out, is it necessary in this method to compute values of $E_j$. 

-17-
Now that I have seen the paper [4] in translation, let me describe it a little more fully. This paper proves the convergence of a process similar to but more general than my own polynomial algorithm. Specifically, the Russian authors consider a class \( \Omega \) of functions \( \Delta \) which can be thought of as a generalization of \( \{ P - f \} \) for a given continuous \( f \), an all polynomials \( P \) of degree at most \( m \). The sequence \( \Delta_0, \Delta_1, \ldots \) is constructed by equating the values of \( \Delta_k \) to \((-1)^i \gamma_k \) at any \( n + 2 \) points satisfying certain properties which are more general than those which I required.

The proofs in [4] are quite different in execution than mine, but the general similarity of direction is readily apparent.

It appears that the method of [4] does not apply to the class of rational approximations which I discussed.


Errata

A minus sign prefixed to a line number indicates the count from the bottom of the page.

Page  Line
1 (Title) delete (U)
1 -1 for September 1, 1960 read 25 April 1960
2 -6 for Tchebyshev read Tschebyscheff
3 -2 for all v read v (three places)
3 -3 for both v read v
3 -6 for $X_1$ read $x_1$
3 -8 for exists read exist
4 -9 for extreme read extrema
4 -1 for extreme read extrema
5 -4 before $F_{m,n}$ insert $£$
5 -7 for $(-1)^i$ read $(-1)^i$
6 -11 for or degree $m-\mu$ read or of degree $m-\mu$
7 -7 delete q
8 -2 for $x_2$ read $x_1$
8 -3 for $x_1$ read $x_1$
8 -1 for of $y \in E$ read or $y \in E$
9 sketch for all $\mu$ read $u$; for 90 read $a_0$. Also, for the higher plane curve sketched with the dashed line, please imagine a parabola with vertex at about $(1, -1/3)$ and passing through $(0, 1/3)$ and $(5/2, 2/3)$.
11 -7 for (N) read N
11 -6 for $E_j(x_i)$ read $|E_j(x_i)|$
11 -5 for $V_j$ read $v_i$
11 -1 for $v_i$ read $v'_i$
12 -1 for $|y_{j+1}$ read $|v_{j+1}$
12 -2 for $v_i$ read $v'_i$
12 -13 for $E_l$ read $E^*$
<table>
<thead>
<tr>
<th>Page</th>
<th>Line</th>
<th>Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>-6</td>
<td>for (y*) read</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>for founded read be</td>
</tr>
<tr>
<td>13</td>
<td>-4</td>
<td>for a j read a j</td>
</tr>
<tr>
<td>13</td>
<td>-2</td>
<td>for matric read met</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>for matric read met</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
<td>for in 6 for read in</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>for For read for</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>for f an all read f and</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>for Review of [3] rep</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>for PUIKER read PC</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>for Process or read</td>
</tr>
</tbody>
</table>