APPROXIMATIONS WITH LEAST MAXIMUM ERROR

R. G. SELFRIDGE

1. **Introduction.** This paper is concerned with the problem of finding best approximations to a given function, where 'best' is defined in a particular fashion, and the approximation is taken from a given class of functions. The approximated function need only be bounded and single-valued; the range of the independent variable can be any closed interval, where the infinite interval is closed at infinity.

This problem was originally investigated from the point of view of obtaining approximations to be used on automatic computing machines in place of functions with slowly converging Taylor series. The result is also of value in determining formulae for empirically determined functions.

This problem has been studied by many people. The present paper extends previous results, and gives an indication of how these approximations can be obtained. Unfortunately, only a sufficient condition has been given, and the convergence of the iteration method given has not been proved.

2. **Theorem.** Let us suppose that a function $f(x)$ is to be approximated over the closed interval $[a, b]$, and that $m \leq f(x) \leq M$ in that interval; $f(x)$ need not be continuous but must be single-valued in $[a, b]$. In practice $f(x)$ will usually have a Taylor series. Take G to be a class of functions, called the class of approximating functions, from which an element $k(x)$ is to be chosen to approximate $f(x)$. From here on it is to be understood that all results will be relative to the interval $[a, b]$ and the particular class G.

DEFINITION 1. If $h(x)$ is continuous and $h(x) \neq 0$, define

$$E_k = \max_{a \leq x \leq b} \frac{|f(x) - k(x)|}{|h(x)|}$$

for all $k(x)$ in G.

In practice, $h(x)$ will be equal to 1 or equal to $f(x)$, to yield actual error or relative error, between $f(x)$ and $k(x)$.

Received February 2, 1952.
Pacific J. Math. 3 (1953), 247-255

DEFINITION 2. The function $g(x) \in G$ is a *best approximation* to $f(x)$, relative to $h(x)$, if $E_g \leq E_k$ for all $k(x)$ in G.

DEFINITION 3. Set

$$e(x) = \frac{f(x) - g(x)}{h(x)}.$$

The *extrema* of $e(x)$ are the relative maxima and minima of $e(x)$. The *absolute extrema* are the $e_j = e(x_j)$ for which $|e(x)| \leq |e_j|$. It should be pointed out that de/dx need not be zero for extrema at the points a, b of the closed interval $[a, b]$.

Now consider G to be the class of functions of the form

$$Q\left[K(x) \sum_{i=0}^{n} a_i x^{p_i}\right]$$

for some fixed n, where $Q(y)$, $K(y)$ are fixed functions, and p_i $(i = 0, 1, \cdots, n)$ is a sequence of positive integers such that $0 = p_0 < p_1 < \cdots < p_n$. It is assumed that $Q(y)$ has an inverse $P(y)$ with a continuous first derivative $P'(y)$, and that $P'(y) \neq 0$ in the interval $[m - HE_g, M + HE_g]$, where H is the maximum absolute value of $h(x)$ as used in Definition 1.

Thus G is a class of functions dependent on the $n + 1$ coefficients a_i, and is, in a sense, a class of generalized polynomials. This choice of G will cover many of the practical cases of approximation with the exception of rational approximations with free coefficients in both numerator and denominator. The requirements of unisolvence used in the approximations of Motzkin [2] include the condition that the difference of two approximations have n or fewer roots in $[a, b]$. This restriction and similar ones that lead to Tchebycheff or Descartes approximations, as used by Bernstein [1], will not permit such examples as powers of polynomials, or polynomials of degree greater than n if $[a, b]$ includes the origin, or products of polynomials and functions that have roots in $[a, b]$. The simplest case of approximation by a polynomial of degree n has been handled by many methods, including the foregoing and others, such as that of de la Vallee Poussin [3]. The results of Motzkin [2] will handle many rational approximations that this paper cannot, and furthermore they supply a necessary and sufficient condition.

LEMMA. *If* $0 \leq x_0 < x_1 < \cdots < x_{n+1}$, *then*

$$A_k^n(x) \equiv A(x_0, x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_{n+1}) \equiv \begin{vmatrix} 1 & x_0^{p_1} & \cdots & x_0^{p_n} \\ 1 & x_1^{p_1} & \cdots & x_1^{p_n} \\ \cdot & \cdot & \cdots & \cdot \\ 1 & x_{k-1}^{p_1} & \cdots & x_{k-1}^{p_n} \\ 1 & x_{k+1}^{p_1} & \cdots & x_{k+1}^{p_n} \\ \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n+1}^{p_1} & \cdots & x_{n+1}^{p_n} \end{vmatrix}$$

is positive definite.

Proof. Clearly this is true for $n = 1$. Suppose it is true for all $m < n$. Then expand $A_k^n(x)$ in terms of x_0. By the induction hypothesis, every n by n minor is positive. Thus

$$A_k^n(x) = B(x_0) = B_0 - B_1 x_0^{p_1} + \cdots + (-1)^n B_n x_0^{p_n},$$

where each $B_i > 0$. By Descartes' rule of signs, the equation $B(x_0) = 0$ has n positive roots, namely $x_0 = x_i$ for $i = 1, 2, \cdots, k-1, k+1, \cdots, n+1$. Also, by the induction hypothesis, B_0 is positive; so we have

$$A_k^n(x) = R(x) \prod_{i \neq k} (x_i - x_0),$$

where $R(x)$ is positive. Since $x_0 < x_i$, $A_k^n(x)$ is positive. The induction holds and the lemma is proved.

THEOREM. *A function $g(x)$ in G, where G is as above, is a best approximation to $f(x)$ in the closed interval $[a, b]$ if the following conditions are satisfied:*

1) *The function $e(x)$ has $n + 2$ absolute extrema $e_j = e(x_j)$ such that $x_j < x_{j+1}$, and $e_j = t_j e_0$, with $t_j = \pm 1$.*

2) *The values $(-1)^j t_j h(x_j) A_j^n / K(x_j)$ are all nonzero, and are all positive or all negative.*

Proof. Consider the $n + 2$ points x_j, where $e_j = e(x_j)$ is an absolute extremum. Let us assume that e_0 is positive. Take any other approximation that is as good as or better than $g(x)$; that is, $E_k \leq E_g$. Set

$$e' = \frac{f(x) - k(x)}{h(x)},$$

and let $e_j' = e'(x_j)$. Then since $E_k \leq E_g$ one must have $e_j' = \theta_j e_j$, where $|\theta_j| \leq 1$.

Consider now the $n+2$ equations in the $n+1$ unknowns a_i,

$$(1)\quad f(x_j) - Q\left[K(x_j) \sum_{i=0}^{n} a_i x_j^{p_i}\right] = h(x_j)\, e(x_j) \qquad (j = 0, 1, \cdots, n+1).$$

Rewritten in simpler notation, (1) is

$$f_j - Q\left(K_j \sum_{i=0}^{n} a_i x_j^{p_i}\right) = h_j e_j,$$

or

$$(2)\qquad \sum_{i=0}^{n} a_i x_j^{p_i} = P(f_j - h_j e_j)/K_j \qquad (j = 0, 1, \cdots, n+1).$$

This is possible since Q is assumed to have an inverse; and by hypothesis $K_j \neq 0$.

For a given set of e_j's there is a solution to (2) if and only if the determinant of the coefficients is zero:

$$(3)\quad H(e) = \begin{vmatrix} 1 & x_0^{p_1} & \cdots & x_0^{p_n} & P(f_0 - h_0 e_0)/K_0 \\ 1 & x_1^{p_1} & \cdots & x_1^{p_n} & P(f_1 - h_1 e_1)/K_1 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 1 & x_{n+1}^{p_1} & \cdots & x_{n+1}^{p_n} & P(f_{n+1} - h_{n+1} e_{n+1})/K_{n+1} \end{vmatrix} = 0.$$

Also, since by hypothesis $A_j^n \neq 0$, if $H(e) = 0$ there is a unique solution. The two sets e_j and e_j' must both satisfy (3); that is, $H(e) = H(e') = 0$. Now set

$$e_j - e_j' = \delta_j = e_j(1 - \theta_j).$$

By the mean-value theorem for several variables, we have

$$H(e) - H(e') = \sum_{j=0}^{n+1} (e_j - e_j') \frac{\partial H}{\partial e_j}, \tag{4}$$

where each of the partial derivatives is evaluated at some intermediate point; (4) may be written as

$$H(e) - H(e') = 0 = \sum_{j=0}^{n+1} \delta_j \frac{\partial H}{\partial e_j}. \tag{5}$$

The partial derivative $\partial H/\partial e_j$ is given by

$$\frac{\partial H}{\partial e_j} = \begin{vmatrix} 1 & x_0^{p_1} & \cdots & x_0^{p_n} & P_0/K_0 \\ \cdot & \cdot & \cdots & \cdot & \cdots \\ 1 & x_{j-1}^{p_1} & \cdots & x_{j-1}^{p_n} & P_{j-1}/K_{j-1} \\ 0 & 0 & \cdots & 0 & -h_j P_j'/K_j \\ 1 & x_{j+1}^{p_1} & \cdots & x_{j+1}^{p_n} & P_{j+1}/K_{j+1} \\ \cdot & \cdot & \cdots & \cdot & \cdots \\ 1 & x_{n+1}^{p_1} & \cdots & x_{n+1}^{p_n} & P_{n+1}/K_{n+1} \end{vmatrix},$$

where P_i and P_j' are evaluated at some intermediate point.

This may be written, by expansion on the jth row, in the form

$$\frac{\partial H}{\partial e_j} = (-1)^{n+j+1} P'(f_j - h_j e_j + h_j \phi_j \delta_j) h_j A_j^n/K_j, \tag{6}$$

where $0 \leq \phi_i \leq 1$.

Thus $\partial H/\partial e_j$ is dependent only on the different x_i, and on h_j, f_j; it is not dependent on e_i, $i \neq j$.

Now we have

$$-h_j e_j + h_j \phi_j \delta_j = -h_j e_j[1 - \phi_j(1 - \theta_j)] = -h_j e_j \zeta_j,$$

where $|\zeta_j| \leq 1$ since $0 \leq \phi_j \leq 1$ and $|\theta_j| \leq 1$.

Therefore

$$P'(f_j - h_j e_j + h_j \phi_j \delta_j) = P'(f_j - h_j e_j \zeta_j),$$

and we have

$$\frac{\partial H}{\partial e_j} = (-1)^{n+j+1} P'(f_j - h_j e_j \zeta_j) h_j A_j^n / K_j, \tag{7}$$

or, rewriting (5),

$$\sum_{j=0}^{n+1} \delta_j \frac{\partial H}{\partial e_j} = \sum_{j=0}^{n+1} (-1)^{n+j+1} P'(f_j - h_j e_j \zeta_j) h_j A_j^n \delta_j / K_j = 0. \tag{8}$$

Now by hypothesis $e_j = t_j e_0$, and we set $\delta_j = e_j(1 - \theta_j)$. Then (8) becomes

$$\sum_{j=0}^{n+1} (-1)^j P'(f_j - h_j e_j \zeta_j) h_j A_j^n e_0 t_j (1 - \theta_j)/K_j = 0. \tag{9}$$

Now $P'(f_j - h_j e_j \zeta_j)$ is always positive, or always negative, by assumption on $Q(x)$; hence, by the second condition of the hypothesis, (9) may be written as

$$\sum_{j=0}^{n+1} e_0 T_j (1 - \theta_j) = 0, \tag{10}$$

where T_j is always positive.

Equation (10) can be satisfied only if $e_0 = 0$, in which case $f(x) = g(x)$, or $\theta_j = 1$; then $k(x) = g(x)$ at the points x_j, and thus $k(x) = g(x)$ for x in $[a, b]$.

Therefore $g(x)$ is a best approximation, and is unique.

COROLLARY 1. *If the origin is not contained in* $[a, b]$ *and* $K(x)$ *and* $h(x)$ *do not change sign in* $[a, b]$, *then, by Lemma* 1, $g(x)$ *is a best approximation if* $e(x)$ *has* $n + 2$ *absolute extrema such that* $e_j = e_0(-1)^j$ *and* $x_j < x_{j+1}$.

COROLLARY 2. *If not all* $A_j^n \neq 0$, *then* $g(x)$ *is still a best approximation provided that at least one* $A_j^n \neq 0$; *but* $g(x)$ *is not necessarily unique.*

This follows since (9) still holds; and (10) still holds but with $T_j \geq 0$ and at least one $T_j \neq 0$. Thus $k(x)$ cannot be better than $g(x)$, but need not be

identical with $g(k)$.

COROLLARY 3. *If* $P'(y)$ *does change sign in*

$$m - HE_g \leq y \leq M + HE_g,$$

then the theorem still holds, but with condition 2) *replaced by*:

2′) *The value* $(-1)^j t_j h_j A_j^n P'(y_j)/K_j$ *is nonzero and positive* (*negative*) *for*

$$f_j - |h_j e_0| \leq y_j \leq |h_j e_0| + f_j \qquad (j = 0, 1, \cdots, n+1).$$

This follows from (9) since it is sufficient that $P'(f_j - h_j e_j \zeta_j)$ not change sign for any ζ_j.

3. **Approximations.** In practice the class G usually has simple functions $Q(x)$, $K(x)$. The function $K(x)$ is chosen to remove some awkward point of $f(x)$, such as a point with an infinite derivative, as occurs for example in $\sin^{-1}x$, while the function $Q(x)$ is taken to be a function like x, x^2 (this requires care, because there is no unique inverse), or $1/x$. The requirements of the theorem force the approximations to be continuous if $K(x)$ is continuous.

The procedure for finding $g(x)$ is to guess some initial value of e_0, to guess a set of points x_i, and to set up $n+1$ equations in order to find the approximation that will go through the $n+1$ points x_i, $f_i - t_i e_0$. This requires using $n+2$ points to compute the t_i. If the error curve is now plotted, then $e(x)$ will have $n+2$ absolute extrema with values e_j' at the points x_j', and the new values will not differ too greatly from the original x_i. That is, $e_j' \approx t_j e_0$ and $x_j \approx x_j'$. Now by averaging, or solution of (3), we find a new e_0 for the points x_j', and repeat the process. This process will usually converge to the desired $g(x)$, but a criterion for convergence has not yet been found. Experience has shown that if the origin is not in $[a, b]$, then convergence is rapid, even from a poor first approximation, but that if the origin is in $[a, b]$ it will be necessary to solve (3) on each iteration at first, and even then the iteration may not converge for a poor first approximation.

Further difficulty can be experienced in that, if there are more than $n+2$ absolute extrema, one choice of $n+2$ may not show that an approximation is best, while another choice will show that the approximation is best. It is therefore necessary to apply the test to all possible sets of $n+2$ absolute extrema to show that a given case is best.

As an example of the iteration process, consider the problem of approximating $\sin x$ in the range $[0, 1]$ by a form $ax + bx^3$. This form satisfies the conditions of Corollary 1, so that the iteration is fairly simple. The functions are considered here in x-intervals of length .02, and the computations have been rounded off. Suppose that at some intermediate stage one has:

$$a = .99987,\ b = -.159403.$$

This yields an error curve with the extrema

$$e_1 = .000006,\ e_2 = -.001,\ e_3 = .001$$

at the points

$$x_1 = .08,\ x_2 = .76,\ x_3 = 1.0.$$

By averaging, or a solution of (3), one finds a new $e_0' = .0007$, and one now finds a new approximation with an error curve that goes through the points

$$(.76, -.0007),\ (1.0, .0007).$$

By a solution of (3), the new error curve could also be made to go through the correct value for x_1, but the extra computation to get this value is frequently not worth the effort. The new curve has

$$a = .99850,\ b = -.157731,$$

with extrema at the points

$$(.24, .00024),\ (.76, -.0007),\ (1.0, .0007).$$

Averaging these extrema yields $e_0 = .0005$, and the approximation with error curve going through the points

$$(.76, -.0005),\ (1.0, .0005)$$

has

$$a = .997605,\ b = -.156634,$$

and extrema at the points

$$(.32, .00046),\ (.8, -.00053),\ (1.0, .0005).$$

Repetition of this process finally stabilizes with the absolute extrema at

.3, .8, 1.0, and we have

$$\sin x \approx .997491x - .1565191x^3 ,$$

with the maximum error less than 5×10^{-4}.

Further refinement is of course possible, by carrying more places in the computation, or by taking more points in [0, 1]; but from the point of view of computational use the additional work is not justified.

The choice of form to be used in an approximation can be decided at present only by trial, as is also the case in all the different ways of short-cutting the iteration.

References

1. S. Bernstein, *Lecons sur les propriétés extrémales, et la meilleure approximation des fonctions analytiques d'une variable réelle*, Gauthier-Villars, Paris, 1926.

2. Th. Motzkin, *Approximation by curves of a unisolvent family*, Bull. Amer. Math. Soc. 55 (1949), 789-793.

3. Ch. J. de la Vallée Poussin, *Sur les polynomes d'approximation et la représentation approchée d'un angle*, Acad. Roy. Belgique. Bull. Cl. Sci. (1910), 808-844.

Naval Ordnance Test Station
China Lake, California