Minimax Approximations Subject to a Constraint

By C. T. Fike and P. H. Sterbenz

Abstract. A class of approximation problems is considered in which a continuous, positive function $\phi(x)$ is approximated by a rational function satisfying some identity. It is proved under certain hypotheses that there is a unique rational approximation satisfying the constraint and yielding minimax relative error and that the corresponding relative-error function does not have an equal-ripple graph. This approximation is, moreover, just the rational approximation to $\phi(x)$ yielding minimax logarithmic error. This approximation, in turn, is just a constant multiple of the rational approximation to $\phi(x)$ yielding minimax relative error but not necessarily satisfying the constraint.

1. Introduction. Various authors have investigated approximation problems in which the approximation $f(x)$ is required to satisfy some functional constraint. For example, Cody and Ralston [1] investigated the problem of finding a rational function $f(x)$ with numerator and denominator of degree $N$ such that $f(x)$ satisfies the constraint

$$f(x) = 1/f(-x)$$

and minimizes the maximum relative error

$$\max_{[-a,a]} \left| \frac{f(x) - e^x}{e^x} \right|.$$  

In this paper, we consider a class of approximation problems including the Cody-Ralston problem and similar problems that have arisen in other contexts. We show that for a problem in this class there is a unique approximation optimal in the sense that it yields minimax relative error, and we characterize this solution.

2. Relative and Logarithmic Error. Suppose that we want to find a polynomial or rational approximation for a function $\phi(x)$ on an interval $I: a \leq x \leq b$, where $\phi(x)$ is continuous and does not vanish in $I$. Then, we may assume that $\phi(x)$ is positive for $x$ in $I$.

Let $V$ be a set of admissible functions. Here $V$ will be either the set of all polynomials of degree $\leq M$ or else the set $V$ will be the set of all rational functions $p(x)/q(x)$ where $p(x)$ and $q(x)$ are relatively prime polynomials of degree $\leq M$ and $\leq N$, respectively, and $q(x)$ does not vanish for $x$ in $I$. We shall refer to such functions $p(x)/q(x)$ as $(M, N)$ rational functions.

For $f(x)$ in $V$, we set

$$R(x) = \frac{f(x) - \phi(x)}{\phi(x)}.$$  

Received November 25, 1969, revised November 2, 1970.

AMS 1969 subject classifications. Primary 41A50, 41A17, 41A40; Secondary 6520, 6525.

Key words and phrases. Rational approximation, polynomial approximation, best approximation, constrained approximation, exponential function, starting approximation for square root.

Copyright © 1971, American Mathematical Society.

295
and let \( \mu \) denote the maximum of \(|R(x)|\) for \( x \) in \( I \). There is a unique function \( f^*(x) \) in \( V \) which minimizes \( \mu \) for all \( f(x) \) in \( V \). We let \( \mu^* \) denote the value of \( \mu \) for \( f^*(x) \).

Let \( W \) be the set of all \( f(x) \) in \( V \) for which \( f(x) > 0 \) for all \( x \) in \( I \). Let \( c \) be the minimum of \( \varphi(x) \) for \( x \) in \( I \). Then the function \( f(x) = c/2 \) is in \( W \), and for this function we have \( \mu < 1 \). But any function which is in \( V - W \) will yield \( \mu \geq 1 \), so \( f^*(x) \) is in \( W \).

For \( f(x) \) in \( W \), we may consider the logarithmic error

\[
\delta(x) = \log_e \frac{f(x)}{\varphi(x)}.
\]

We shall use \( \lambda \) to designate the maximum of \(|\delta(x)|\) for \( x \) in \( I \). Thus, with any function \( f(x) \), we associate values of \( \lambda \) and \( \mu \). Clearly,

\[
R(x) = e^{\delta(x)} - 1.
\]

Instead of trying to find \( f^*(x) \), it is sometimes convenient to try to find a function \( f(x) \) in \( W \) which minimizes \( \lambda \).

In [2], we proved the following theorem for the special case in which \( \varphi(x) = \sqrt{x} \). However, the proof given there is valid for any positive continuous function \( \varphi(x) \), so it will not be repeated here.

**Theorem 1.** There is a unique function \( \tilde{f}(x) \) in \( W \) which minimizes the maximum of \(|\delta(x)|\) on \( I \) for all \( f(x) \) in \( W \). If \( \tilde{\lambda} \) is the value of \( \lambda \) for \( \tilde{f}(x) \), we have

\[
\tilde{\lambda} = \text{arc tanh} \ \mu^*.
\]

\( \tilde{f}(x) \) is characterized by the fact that it produces an equal-ripple \( \delta(x) \), and it is related to \( f^*(x) \) by

\[
\tilde{f}(x) = f^*(x)/(1 - (\mu^*)^2)^{1/2},
\]

\[
f^*(x) = \tilde{f}(x)/\cosh \tilde{\lambda}.
\]

### 3. Constraints.

In addition to the two related problems of finding \( f^*(x) \) and \( \tilde{f}(x) \), there are some cases in which it is desirable to consider a third problem in which \( f(x) \) is required to satisfy an identity satisfied by \( \varphi(x) \). Three examples are:

1. Find the best \((N, N)\) rational approximation \( f(x) \) for \( e^x \) on \(-\alpha \leq x \leq \alpha\) such that \( f(-x) = 1/f(x) \).

2. For \( 0 < \alpha < 1 \), find the best \((N, N)\) rational approximation \( f(x) \) for \( \sqrt{x} \) on \( \alpha \leq x \leq 1/\alpha \) such that \( f(1/x) = 1/f(x) \).

3. For \( N > 0 \) and \( 0 < \alpha < 1 \), find the best \((N + 1, N)\) rational approximation \( f(x) \) for \( \sqrt{x} \) on \( \alpha \leq x \leq 1/\alpha \) such that \( xf(1/x) = f(x) \).

In each case, by the best approximation, we mean the one which minimizes \( \mu \) subject to the constraint. An approximation of the first type is found by Cody and Ralston in [1] and by Kahan in [3]. Maehly studied an approximation of the second type. See the appendix of [4]. In [4], Cody finds an approximation of the third type. These constraints often simplify the problem of finding the best approximation by reducing the number of coefficients.

In each case, we have a constraint \( C \). Let \( U \) be the set of all functions \( f(x) \) in \( V \) which satisfy the constraint \( C \). We shall require that the set \( U \) have the following properties:

\[(a) \ \tilde{f}(x) \text{ is in } U.\]
(b) If \( f(x) \) is in \( U \cap W \), then for any \( x \) in \( I \) there is a point \( y \) in \( I \) such that \( \delta(y) = -\delta(x) \).

(c) For any \( f(x) \) in \( U - W \) there is a \( g(x) \) in \( U \cap W \) which has a smaller \( \mu \) than \( f(x) \) does.

We first show that for each of the three examples considered above, \( U \) satisfies these properties. That \( \tilde{f}(x) \) is in \( U \) follows from the uniqueness of \( \tilde{f}(x) \), since otherwise we would have another function in \( W \) with the same value of \( \lambda \), namely \( 1/f(-x) \) in (1), \( f(1/x) \) in (2), and \( xf(1/x) \) in (3). For property (b) of \( U \), we use \( y = -x \) in (1) and \( y = 1/x \) in (2) and (3). For property (c) of \( U \), we first observe that our definition of \( V \) implies that every function \( f(x) \) in \( V \) is bounded on \( I \). For examples (1) and (2), this implies that \( f(x) \) cannot vanish in the interval \( I \), so if \( f(x) \) is in \( U - W \), we take \( g(x) = -f(x) \). In the third example, we may always take \( g(x) = \epsilon + cx \), where \( \epsilon \) is a small positive constant such that the maximum of \( g(x) \) is less than the minimum of \( \varphi(x) \) for \( x \) in \( I \).

We now address the problem of finding \( f(x) \) in \( U \) which minimizes \( \mu \). Because of property (c), we need consider only functions in \( U \cap W \). But for any function \( f(x) \) in \( U \cap W \), we have, by (1), \( e^{\lambda} - 1 \geq R(x) \geq e^{-\lambda} - 1 \), and since \( \delta(x) \) is continuous on \( I \) there is a point \( x \) in \( I \) with \( |\delta(x)| = \lambda \). But by property (b), there is a point \( y \) in \( I \) with \( \delta(y) = -\delta(x) \), so \( R(x) \) assumes both the values \( e^{\lambda} - 1 \) and \( e^{-\lambda} - 1 \) in \( I \).

Then, for \( f(x) \) we have

\[
\mu = e^{\lambda} - 1.
\]

Since \( \tilde{f}(x) \) minimizes \( \lambda \) for all \( f(x) \) in \( W \), we have \( \lambda \geq \tilde{\lambda} \), and therefore (2) implies \( \mu \geq e^{\tilde{\lambda}} - 1 \). By property (a), \( \tilde{f}(x) \) is in \( U \cap W \). Then, using \( \bar{\mu} \) to denote the value of \( \mu \) for \( \tilde{f}(x) \), we have, from (2), \( \bar{\mu} = e^{\tilde{\lambda}} - 1 \). Then, \( \tilde{f}(x) \) minimizes \( \mu \) for all \( f(x) \) in \( U \). If \( g(x) \) is any function in \( U \) with

\[
\mu = e^{\lambda} - 1,
\]

then (2) and (3) imply that \( \lambda = \tilde{\lambda} \), so the uniqueness of the function minimizing the maximum of \( |\delta(x)| \) implies that \( g(x) = \tilde{f}(x) \). We have proved:

**Theorem 2.** \( \tilde{f}(x) \) is the unique function in \( U \) which minimizes the maximum of \( |R(x)| \) for all \( f(x) \) in \( U \). For \( \tilde{f}(x) \), we have

\[
e^{-\lambda} - 1 \leq R(x) \leq e^{\lambda} - 1 \quad \text{and} \quad \bar{\mu} = e^{\tilde{\lambda}} - 1.
\]

The relation between the solution \( \tilde{f}(x) \) of the constrained problem and the solution \( f^*(x) \) of the unconstrained problem is given in Theorem 1.

4. Comments. Since \( \tilde{f}(x) \) produces an equal-ripple \( \delta(x) \), it produces an \( R(x) \) which has the correct number of alternating sign extrema but which is not equal-ripple because the maximum is larger than the absolute value of the minimum. Thus, with constraints of this sort, the best-fit problem has a solution which does not produce an equal-ripple error curve.

For the first example, approximating \( e^x \), we would usually select \( V \) so that the approximation \( f^*(x) \) is accurate to better than word length. Since

\[
(1 - (\mu^*)^2)^{1/2} \approx 1 - \frac{1}{2}(\mu^*)^2,
\]

this means that \( f^*(x) \) and \( \tilde{f}(x) \) agree to more than twice word length, and so do...
Thus, we will be equally satisfied with either \( f^*(x) \) or \( \tilde{f}(x) \). Since the constraint reduces the number of coefficients, it may be easier to consider the constrained problem.

For \( \sqrt{x} \), we usually look for a starting approximation, and then use Newton’s method. In this case, \( f^*(x) \) and \( \tilde{f}(x) \) may be noticeably different, since the approximation is not very accurate. But we showed in [2] that \( \tilde{f}(x) \) minimizes the maximum relative error after one or more iterations, so we would prefer to have \( \tilde{f}(x) \) instead of \( f^*(x) \). Then the constraint may be used to simplify the computation as in [4].

IBM Systems Research Institute
New York, New York 10017