MATHEMATICS OF COMPUTATION, VOLUME 25, NUMBER 114, APRIL, 1971

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 $\varphi(x)$ 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 $\varphi(x)$ yielding minimax logarithmic error. This approximation, in turn, is just a constant multiple of the rational approximation to $\varphi(x)$ 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

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

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 $\varphi(x)$ on an interval I: $a \leqq x \leqq b$, where $\varphi(x)$ is continuous and does not vanish in I. Then, we may assume that $\varphi(x)$ 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 $\leqq M$ 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 $\leqq M$ and $\leqq N$, 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

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

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

AMS 1969 *subject classifications*. Primary 4115, 4117, 4140; Secondary 6520, 6525.

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

and let μ denote the maximum of $|R(x)|$ for x in I. There is a unique function $f^*(x)$ in V which minimizes μ for all $f(x)$ in V. We let μ^* denote the value of μ 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 $\mu \geqq 1$, 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 λ to designate the maximum of $|\delta(x)|$ for x in I. Thus, with any function $f(x)$, we associate values of λ and μ. Clearly,

$$R(x) = e^{\delta(x)} - 1. \tag{1}$$

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

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 $\bar{f}(x)$ in W which minimizes the maximum of $|\delta(x)|$ on I for all $f(x)$ in W. If $\bar{\lambda}$ is the value of λ for $\bar{f}(x)$, we have*

$$\bar{\lambda} = \operatorname{arc\,tanh} \mu^*.$$

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

$$\bar{f}(x) = f^*(x)/(1 - (\mu^*)^2)^{1/2},$$

$$f^*(x) = \bar{f}(x)/\cosh \bar{\lambda}.$$

3. Constraints. In addition to the two related problems of finding $f^*(x)$ and $\bar{f}(x)$, 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 $-\alpha \leqq x \leqq \alpha$ 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 $\alpha \leqq x \leqq 1/\alpha$ 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 $\alpha \leqq x \leqq 1/\alpha$ such that $xf(1/x) = f(x)$.

In each case, by the best approximation, we mean the one which minimizes μ 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) $\bar{f}(x)$ 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 μ than $f(x)$ does.

We first show that for each of the three examples considered above, U satisfies these properties. That $\bar{f}(x)$ is in U follows from the uniqueness of $\bar{f}(x)$, since otherwise we would have another function in W with the same value of λ, 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 $g(x) = \epsilon + \epsilon x$, where ϵ 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 μ. 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), $e^{\lambda} - 1 \geqq R(x) \geqq e^{-\lambda} - 1$, 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

$$\mu = e^{\lambda} - 1. \tag{2}$$

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

$$\mu = e^{\bar{\lambda}} - 1, \tag{3}$$

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

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

$$e^{-\bar{\lambda}} - 1 \leqq \bar{R}(x) \leqq e^{\bar{\lambda}} - 1 \quad \textit{and} \quad \bar{\mu} = e^{\bar{\lambda}} - 1.$$

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

4. Comments. Since $\bar{f}(x)$ 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

$$(1 - (\mu^*)^2)^{1/2} \approx 1 - \tfrac{1}{2}(\mu^*)^2,$$

this means that $f^*(x)$ and $\bar{f}(x)$ agree to more than twice word length, and so do

$e^{\bar{\lambda}} - 1$ and $|e^{-\bar{\lambda}} - 1|$. Thus, we will be equally satisfied with either $f^*(x)$ or $\bar{f}(x)$. 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 $\bar{f}(x)$ may be noticeably different, since the approximation is not very accurate. But we showed in [2] that $\bar{f}(x)$ minimizes the maximum relative error after one or more iterations, so we would prefer to have $\bar{f}(x)$ 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

1. W. J. Cody & Anthony Ralston, "A note on computing approximations to the exponential function," *Comm. ACM*, v. 10, 1967, pp. 53–55.
2. P. H. Sterbenz & C. T. Fike, "Optimal starting approximations for Newton's method," *Math. Comp.*, v. 23, 1969, pp. 313–318. MR **39** #6511.
3. W. Kahan, "Library tape functions EXP, TWOXP, and .XPXP.," *Programmers' Reference Manual*, University of Toronto, 1966. (Mimeographed.)
4. W. J. Cody, "Double-precision square root for the CDC-3600," *Comm. ACM*, v. 7, 1964, pp. 715–718.