Necessary conditions for a minimax approximation

By A. R. Curtis and M. J. D. Powell*

$\phi(x,\lambda_1^*,\lambda_2^*,\ldots,\lambda_n^*)$ is a minimax approximation to $f(x)$ if the values $\lambda_i = \lambda_i^*$, $i = 1$ to n of the parameters are such that the maximum value (over x) of $|f(x) - \phi(x,\lambda)|$ is minimized. Rice (1960) has established a number of conditions characterizing minimax approximations, but these apply only to a limited class of approximating functions, $\phi(x,\lambda)$. A simple example serves to illustrate that more general conditions are required—some are derived in this paper.

1. Introduction

Rice (1964) has provided a valuable survey of the current theory of minimax approximations in the event that the approximating functions $\phi(x, \lambda_1, \lambda_2, \ldots, \lambda_n)$ take the form $\sum_{i=1}^{n} \lambda_i \phi_i(x)$. It is known that if the functions $\phi_i(x)$ form a Chebyshev set, the best approximation to a continuous function, $f(x)$, over a closed interval, $a \leqslant x \leqslant b$, is characterized by the maximum error occurring at $(n + 1)$ points, the sign of the error alternating. Rice (1960) has extended this theorem to include non-linear dependence of $\phi(x, \boldsymbol{\lambda})$ on $\boldsymbol{\lambda}$ by identifying conditions on $\phi(x, \boldsymbol{\lambda})$ which are necessary and sufficient to ensure that the same characterization of best approximations holds for all continuous $f(x)$.

Unfortunately there are many useful choices of $\phi(x, \boldsymbol{\lambda})$ which do not fulfil Rice's conditions, notably rational approximations

$$\phi(x, \boldsymbol{\lambda}) = \left[\sum_{i=1}^{m} \lambda_i x^{i-1}\right] \Big/ \left[1 + \sum_{i=m+1}^{n} \lambda_i x^{i-m}\right].$$

In this case the defect, d, has to be included in Chebyshev's characterization theorem (Achieser, 1956). An example of another exception is provided by the approximation of x^2 by $\lambda_1 x + \lambda_2 e^x$ over $0 \leqslant x \leqslant 2$. This will be considered in some detail. It may be verified that the error function of the approximation

$$x^2 \approx 8{\cdot}465x - 2{\cdot}0239e^x$$

takes its maximum absolute value at the three points $x = 0$, $x = 1{\cdot}1227$ and $x = 2$, the error at these points being $+2{\cdot}0239$, $-2{\cdot}0239$ and $+2{\cdot}0239$, respectively. However, note that because $(e^x - 3{\cdot}5x)$, for example, has the same sign oscillation properties as the maximum error, the approximation can be improved by subtracting a positive multiple of $(e^x - 3{\cdot}5x)$. In fact the best approximation is

$$x^2 \approx 0{\cdot}1842x + 0{\cdot}4186e^x,$$

the maximum absolute error is $0{\cdot}5382$, and this error occurs at just the two points $x = 0{\cdot}4064$ and $x = 2$. Rice's theorems are not applicable because $(e^x - 3{\cdot}5x)$ has two zeros in the range $0 \leqslant x \leqslant 2$, but the simplicity of this example suggests that more general characterization theorems are of importance.

In this paper methods for finding minimax approximations are not considered. It is assumed that the best values of the parameters have been determined, and Theorem 1 states a necessary condition for a best approximation when the maximum error occurs at fewer than $(n + 1)$ points. Theorem 2 relates the signs of the error extrema to the signs of the determinants in the event that the maximum error does occur at $(n + 1)$ points, a best approximation does not necessarily possess the property that the signs of the maximum errors alternate.

2. Notation

The range of x over which the approximation is to apply is called S. L is reserved for the range of the parameters.

The maximum error of an approximation is called $h(\boldsymbol{\lambda})$, so that

$$h(\boldsymbol{\lambda}) = \max_{x \in S} |f(x) - \phi(x, \boldsymbol{\lambda})|. \tag{1}$$

The parameters of a minimax approximation are denoted by $\boldsymbol{\lambda}^*$. $h(\boldsymbol{\lambda}^*)$ is therefore the greatest lower bound of $h(\boldsymbol{\lambda})$—it is abbreviated by h^*.

The values of x at which the maximum error of the approximation $\phi(x, \boldsymbol{\lambda}^*)$ occurs are called $\xi_1, \xi_2, \ldots, \xi_r$. For each ξ_i a number s_i is defined to be ± 1 according to the sign of the error. Therefore

$$f(\xi_i) - \phi(\xi_i, \boldsymbol{\lambda}^*) = s_i h^*; \; i = 1, 2, \ldots, r. \tag{2}$$

In the theorems the derivatives of $\phi(x, \boldsymbol{\lambda})$ with respect to the parameters are required at $\xi_1, \xi_2, \ldots, \xi_r$. They are denoted by

$$D_{ij} = \left[\frac{\partial \phi(\xi_i, \boldsymbol{\lambda})}{\partial \lambda_j}\right]_{\boldsymbol{\lambda} = \boldsymbol{\lambda}^*}. \tag{3}$$

$\boldsymbol{D}$ is the $r \times n$ matrix whose elements are given by (3).

In the event that $r = n + 1$, square matrices of order n, $\Delta_1, \Delta_2, \ldots, \Delta_r$, are defined. Δ_k is the matrix obtained by deleting the kth row of $\boldsymbol{D}$. ρ_k is reserved for the determinant of Δ_k multiplied by $(-1)^k$. An important consequence of this definition is that

$$\sum_{k=1}^{r} \rho_k D_{kj} = 0; \; j = 1, 2, \ldots, n. \tag{4}$$

* *Applied Mathematics Group, Theoretical Physics Division, Atomic Energy Research Establishment, Harwell, Berkshire.*

3. The theorems

The assumptions made in proving the theorems are stated in Section 4. In Section 8 some will be relaxed.

Theorem 1. The rank of $\boldsymbol{D}$ is less than r.

Theorem 2. If $r = n + 1$, the signs of $s_1, s_2, \ldots, s_r$ are all the same as or are all opposite to the signs of $\rho_1, \rho_2, \ldots, \rho_r$.

4. Assumptions made in proving the theorems

For reference purposes the five assumptions made are distinguished by letters. Most important is

(a) A best approximation $\phi(x, \boldsymbol{\lambda}^*)$ exists.

This can be guaranteed if L is closed and bounded, and $\phi(x, \boldsymbol{\lambda})$ is continuous in $\boldsymbol{\lambda}$.

(b) S is composed of a finite number of closed intervals of real numbers and a finite number of discrete real points.

(c) In each closed interval of S, both $f(x)$ and $\phi(x, \boldsymbol{\lambda}^*)$ are continuous functions of x.

(b) and (c) will be relaxed in Section 8.

(d) $\boldsymbol{\lambda}^*$ is an interior point of L.

(d) rules out some cases covered by the comment on (a) above.

(e) In a neighbourhood of $\boldsymbol{\lambda}^*$ and for all $x \varepsilon S$, the derivatives

$$\frac{\partial}{\partial \lambda_i}\phi(x, \boldsymbol{\lambda}), \quad i = 1, 2, \ldots, n,$$

are uniformly continuous functions of x and $\boldsymbol{\lambda}$.

From (d) and (e) it follows that numbers ω and Ω exist such that, for all $\boldsymbol{\lambda}$ satisfying $||\boldsymbol{\lambda} - \boldsymbol{\lambda}^*|| < \omega$ and for all $x \varepsilon S$, $\boldsymbol{\lambda} \varepsilon L$ and

$$\left|\frac{\partial}{\partial \lambda_i}\phi(x, \boldsymbol{\lambda})\right| < \Omega. \tag{5}$$

5. Proof of Theorem 1

Since $\boldsymbol{D}$ is an $r \times n$ matrix, Theorem 1 is true if r exceeds n. Therefore it need only be proved for $r \leqslant n$ and, in this case, it will be shown that assuming $\boldsymbol{D}$ is of rank r leads to a contradiction.

If $\boldsymbol{D}$ is of rank $r \leqslant n$, a vector $\boldsymbol{\lambda}^\dagger$ may be found such that

$$\sum_{j=1}^{n} D_{ij}\lambda_j^\dagger = s_i; \; i = 1, 2, \ldots, r. \tag{6}$$

The contradiction is that the maximum error of the approximation $\phi(x, \boldsymbol{\lambda}^* + \mu\boldsymbol{\lambda}^\dagger)$ is less than h^* for sufficiently small positive values of μ. The error function of this approximation is called

$$e(x, \mu) = f(x) - \phi(x, \boldsymbol{\lambda}^* + \mu\boldsymbol{\lambda}^\dagger). \tag{7}$$

Note that, in virtue of (6),

$$\left[\frac{\partial}{\partial \mu}e(\xi_i, \mu)\right]_{\mu=0} = -s_i; \; i = 1, 2, \ldots, r, \tag{8}$$

so that the error magnitudes at the points ξ_i are initially decreasing functions of μ.

For the proof S is divided into two parts, R_N and T_N. T_N contains $\xi_1, \xi_2, \ldots, \xi_r$ and R_N is closed and bounded. We show that $|e(x, \mu)|$ may be made less than h^* in each part; the method of proof depends on whether x is in R_N or T_N. For the present purpose, N may be any fixed number in the range $0 < N < 1$; it will be given a specific value in the proof of Theorem 2.

x is defined to belong to T_N if all the following three conditions are satisfied:

(i) $|e(x, 0)| > \frac{1}{2}h^*$

(ii) $\left|\left[\frac{\partial}{\partial \mu}e(x, \mu)\right]_{\mu=0}\right| > N$

(iii) the sign of $\left[\frac{\partial}{\partial \mu}e(x, \mu)\right]_{\mu=0}$ is opposite to the sign of $e(x, 0)$.

Otherwise $x \varepsilon R_N$. That $\xi_1, \xi_2, \ldots, \xi_r$ all belong to T_N follows from (2), (7) and (8). That R_N is closed and bounded follows from assumption (b) and the strict inequalities in conditions (i) and (ii) above.

Remembering assumption (c),

$$\max_{x \epsilon R_N} |e(x, 0)| = h_R, \text{ say}, \tag{9}$$

is attained. Therefore

$$h_R < h^*. \tag{10}$$

By (5) and (7), provided $|\mu| < \omega/||\boldsymbol{\lambda}^\dagger||$,

$$\left|\frac{\partial}{\partial \mu}e(x, \mu)\right| < \Omega \sum_{i=1}^{n} |\lambda_i^\dagger| = M, \text{ say}. \tag{11}$$

Hence, provided $|\mu|$ is restricted to be less than both $\omega/||\boldsymbol{\lambda}^\dagger||$ and $(h^* - h_R)/M$, it will follow that

$$\max_{x \epsilon R_N} |e(x, \mu)| < h^*. \tag{12}$$

Because of assumption (e) and by condition (ii) on T_N, a positive number m, not exceeding $\omega/||\boldsymbol{\lambda}^\dagger||$, may be chosen such that, provided $x \varepsilon T_N$ and $|\mu| < m$,

$$\left|\left[\frac{\partial}{\partial \mu}e(x, \mu)\right]\right| > \tfrac{1}{2}N. \tag{13}$$

Where (13) is satisfied $e(x, \mu)$ is a strictly monotonic function of μ, as is $|e(x, \mu)|$ provided that $e(x, \mu)$ remains non-zero. Condition (i) of T_N ensures that $e(x, \mu)$ is non-zero if $|\mu| < \frac{1}{2}h^*/M$. Further, condition (iii) of T_N ensures that, in a neighbourhood of $\mu = 0$, $|e(x, \mu)|$ is a strictly decreasing function of μ. Hence, provided $|\mu| < m$, $|\mu| < \frac{1}{2}h^*/M$ and $\mu > 0$,

$$\max_{x \epsilon T_N} |e(x, \mu)| < h^*. \tag{14}$$

(12) and (14) provide the required contradiction, so it must be concluded that it is impossible to solve (6), and the rank of $\boldsymbol{D}$ is less than r.

6. Proof of Theorem 2

It is proved that the signs of $s_1, s_2, \ldots, s_r$ are all the same as or are all opposite to the signs of $\rho_1, \rho_2, \ldots, \rho_r$ by showing that, for all k, $\rho_k s_k$ has the sign of

$$\sigma = \sum_{i=1}^{r} \rho_i s_i. \tag{15}$$

It will be apparent that, unless $\rho_1 = \rho_2 = \ldots = \rho_r = 0$, σ is non-zero because in fact a stricter result is proved. It is that if both $\rho_k s_k$ and $(\sigma - \rho_k s_k)$ are non-zero, they have the same sign. Should ρ_k be zero, its sign must be interpreted favourably. Should $(\sigma - \rho_k s_k)$ be zero, $\sigma = \rho_k s_k$.

Again the proof depends on a contradiction, so it is supposed that, for some k, $\rho_k s_k$ and $(\sigma - \rho_k s_k)$ are non-zero and have opposite signs. As ρ_k is not zero, Δ_k is non-singular, and $\boldsymbol{\lambda}^\dagger$ is defined by the equations

$$\Delta_k \boldsymbol{\lambda}^\dagger = s, \tag{16}$$

where the transpose of s is

$$(s_1, s_2, \ldots, s_{k-1}, s_{k+1}, \ldots, s_r).$$

$e(x, \mu)$ is defined as in equation (7), but equation (8) becomes

$$\left[\frac{\partial}{\partial \mu} e(\xi_i, \mu)\right]_{\mu=0} = -s_i; \quad i = 1, 2, \ldots, k-1, k+1, \ldots, r$$

$$\left[\frac{\partial}{\partial \mu} e(\xi_k, \mu)\right]_{\mu=0} = -\sum_{j=1}^{n} D_{kj} \lambda_j^\dagger. \tag{17}$$

From equation (17)

$$\rho_k \left[\frac{\partial}{\partial \mu} e(\xi_k, \mu)\right]_{\mu=0} = -\sum_{j=1}^{n} \rho_k D_{kj} \lambda_j^\dagger. \tag{18}$$

Hence, using (4), (15) and (16),

$$\begin{aligned} \rho_k \left[\frac{\partial}{\partial \mu} e(\xi_k, \mu)\right]_{\mu=0} &= \sum_{j=1}^{n} \left\{ \sum_{\substack{i=1 \\ i \neq k}}^{r} \rho_i D_{ij} \right\} \lambda_j^\dagger \\ &= \sum_{\substack{i=1 \\ i \neq k}}^{r} \rho_i \left\{ \sum_{j=1}^{n} D_{ij} \lambda_j^\dagger \right\} \\ &= \sum_{\substack{i=1 \\ i \neq k}}^{r} \rho_i s_i \\ &= \sigma - \rho_k s_k. \end{aligned} \tag{19}$$

Therefore, by the hypothesis on $\rho_k s_k$ and $(\sigma - \rho_k s_k)$, $\left[\frac{\partial}{\partial \mu} e(\xi_k, \mu)\right]_{\mu=0}$ is non-zero, and its sign is opposite to that of s_k. Hence, choosing N to be the smaller of $\frac{1}{2}$ and $\frac{1}{2}\left|\left[\frac{\partial}{\partial \mu} e(\xi_k, \mu)\right]_{\mu=0}\right|$, and defining T_N as before, again $\xi_1, \xi_2, \ldots, \xi_r$ all belong to T_N. Thus a contradiction results so Theorem 2 must hold.

7. An application of the theorems

The two theorems are of use in studying the example of Section 1, namely

$$x^2 \approx \lambda_1 x + \lambda_2 e^x.$$

The approximation $8{\cdot}4656x - 2{\cdot}0239e^x$ has $r = n + 1$ and

$$\xi_1 = 0{\cdot}0000, \xi_2 = 1{\cdot}1227, \xi_3 = 2{\cdot}0000,$$

$$s_1 = +1 \quad , s_2 = -1 \quad , s_3 = +1$$

and $\rho_1 = 2{\cdot}1495, \rho_2 = 2{\cdot}0000, \rho_3 = -1{\cdot}1227.$

Therefore, by Theorem 2, it is not a best approximation.

On the other hand $0{\cdot}1842x + 0{\cdot}4186e^x$ has $r = 2$ so, if it is a best approximation, Theorem 1 should not be trivially satisfied. Since

$$\boldsymbol{D} = \begin{pmatrix} 0{\cdot}4064 & 1{\cdot}5014 \\ 2{\cdot}0000 & 7{\cdot}3891 \end{pmatrix}$$

(within rounding errors), its rank is unity as predicted by the theorem.

8. Extensions to the theorems

It has been stated that assumptions (*b*) and (*c*) of Section 4 may be relaxed. Their only purpose is to prove (10), which was deduced from the fact that R_N is closed and bounded. Otherwise, to prove the theorems, it is necessary to define h_R as the least upper bound of $|e(x, 0)|$ as x ranges over R_N, and to establish that (10) still holds.

Without assumptions (*b*) and (*c*), there may be a set of points of S, $\bar{x}_1, \bar{x}_2, \ldots$ such that

$$\lim_{t \to \infty} |e(\bar{x}_t, 0)| = h^*.$$

Further, an infinite number of members of the set may not belong to T_N because of conditions (ii) and/or (iii). In this case the proof breaks down, and it is necessary to restate the theorems in a way which includes limit points of such sets among $\xi_1, \xi_2, \ldots, \xi_r$.

Such sets necessarily have limit points only if S is finite, so this assumption will be made in place of (*b*) and (*c*).

Suppose $\xi_1, \xi_2, \ldots, \xi_p$ are the points of S at which the maximum error of $\phi(x, \boldsymbol{\lambda}^*)$ is attained; p may be zero. As many points as possible are appended to this set, but the additions must comply with the following rules. Suppose the current set is $\xi_1, \xi_2, \ldots, \xi_{i-1}$. ξ_i is added if there exists a sequence $\bar{x}_{i1}, \bar{x}_{i2}, \ldots$ such that

$$\bar{x}_{ij} \varepsilon S, \quad j = 1, 2, \ldots, \tag{20}$$

$$\lim_{t \to \infty} \bar{x}_{it} = \xi_i \tag{21}$$

and

$$\lim_{t \to \infty} \{f(\bar{x}_{it}) - \phi(\bar{x}_{it}, \boldsymbol{\lambda}^*)\} = s_i h^*, \tag{22}$$

where again $s_i = \pm 1$. Further, there must not exist $j < i$ such that $\xi_i = \xi_j$ *and* $s_i = s_j$. The number of points when no more can be appended is called r.

For the extended set $\xi_1, \xi_2, \ldots, \xi_r$, the theorems do not depend on assumptions (*b*) and (*c*). Note that assumption (*e*) is required to ensure that the matrix elements D_{ij} are well defined. To prove the theorems it is straightforward to show that for each sequence $\bar{x}_{i1}, \bar{x}_{i2}, \ldots$ there exists t_i such that $\bar{x}_{ij}$ belongs to T_N for all $j > t_i$. Should some $\xi_i = \xi_j, s_i \neq s_j$, the theorems are trivial.

As an example consider the best linear approximation to the hypothetical function

$$f(x) = 1 + x, \quad x > 0$$
$$f(x) = \tfrac{1}{2}, \quad x = 0$$
$$f(x) = 0, \quad x < 0 \text{ and irrational}$$
$$f(x) = -x, \quad x < 0 \text{ and rational},$$

where S is composed of the two intervals $-1 \leqslant x \leqslant -\frac{1}{2}$ and $0 < x < 1$. The best approximation is

$$f(x) \approx \phi(x, \boldsymbol{\lambda}^*) = \tfrac{1}{2}x + \tfrac{7}{8}$$

and it has a maximum error of $+\frac{5}{8}$ attained only at $x = -1$. However, as well as $\xi_1 = -1$, the points $\xi_2 = -\frac{1}{2}$ and $\xi_3 = 1$ must be included because

$$\lim_{\substack{x \to -\frac{1}{2} \\ x < -\frac{1}{2} \text{ and irrational}}} \{f(x) - \phi(x, \boldsymbol{\lambda}^*)\} = -\tfrac{5}{8}$$

and

$$\lim_{\substack{x \to 1 \\ x < 1}} \{f(x) - \phi(x, \boldsymbol{\lambda}^*)\} = +\tfrac{5}{8}.$$

Because the approximation is linear, Theorems 1 and 2 state the well-known sign-alternation properties of the maximum error, and this is confirmed by the example.

There are many practical examples in which an approximation is required over an infinite range, so it is of interest to consider imposing no conditions on S. Obviously none are necessary if the infinite range can be transformed to a finite one in a way which preserves the remaining assumptions. Only (*e*) is dependent on such a transformation, so no restrictions need be imposed on the range of the approximation if a change of variables can be found such that S becomes finite and such that

$$\frac{\partial}{\partial \lambda_i} \phi(x, \boldsymbol{\lambda}), \quad i = 1, 2, \ldots, n,$$

are uniformly continuous functions of the new variable over the new range. For this to be the case it is necessary that

$$\lim_{x \to \infty} \frac{\partial}{\partial \lambda_i} \phi(x, \boldsymbol{\lambda}) \tag{23}$$

exist. Even if (23) does not hold, the theorems still apply if there exists a number H and a number $\eta < h^*$ such that

$$|f(x) - \phi(x, \boldsymbol{\lambda}^*)| \leqslant \eta \tag{24}$$

provided that $x \varepsilon S$ and $|x| > H$, because this is the condition that the best approximation over S is also the best approximation over a finite range contained in S.

The last extension may be the most important. It is that the assumption that S is one-dimensional has not been used and, from the beginning, the analysis could have been applied to approximations to functions of several variables.

9. Conclusion

The theorems of this paper serve in two ways which are illustrated by the example of Section 7. First, Theorem 2 may reject an approximation which appears to minimize the maximum error, but does not. Secondly, if by some means an approximation has been obtained whose error attains its maximum value at fewer than $(n + 1)$ points, Theorem 1 may reject it. If it fails to do so, this lends support to the view that a best approximation has been found, without unfortunately being conclusive. A general condition depending on local properties can presumably never be sufficient.

It is hoped that the most useful view has been taken on the conditions that the functions have to satisfy. Usually $\phi(x, \boldsymbol{\lambda})$ is easy to calculate, so seldom are its conditions restrictive. On the other hand, $f(x)$ might be the result of experimental observations or it might not be feasible to calculate it to high accuracy. Further it might only be obtainable for a limited range of values of x, so the fact that it and S are virtually unrestricted is important.

References

ACHIESER, N. I. (1956). *Theory of Approximation*, New York: Frederick Ungar Publishing Co.

RICE, J. R. (1960). "The Characterization of Best Nonlinear Tchebycheff Approximations," *Trans. Amer. Math. Soc.*, Vol. 96, pp. 322–340.

RICE, J. R. (1964). *The Approximation of Functions*, Vol. 1, Addison-Wesley Publishing Co.