ON THE EXISTENCE OF A BEST APPROXIMATION OF ONE DISTRIBUTION FUNCTION BY ANOTHER OF A GIVEN TYPE

By D. L. Burkholder

University of Illinois

1. Introduction and summary. For well over two centuries mathematicians have considered the conditions under which it is possible to obtain a good approximation of one probability distribution by another of a given type. However, the conditions which assure the existence of a best approximation of a given type seem to have been virtually neglected. Because of their intrinsic interest and because of their relevance for an estimation problem which is discussed later, such conditions are examined here with respect to the following example: Suppose that F and G are distribution functions and that an ordered real number pair (a, b), with $a > 0$, is desired such that $F(ax + b)$ is close to $G(x)$ for all real x. Is there a best pair? For instance, is there a pair (a_0, b_0) satisfying

$$(1) \quad \sup_{-\infty<x<\infty} | F(a_0 x + b_0) - G(x) | = \inf_{\substack{0<a<\infty \\ -\infty<b<\infty}} \sup_{-\infty<x<\infty} | F(ax + b) - G(x) | \ ?$$

In this note we give an example in which a pair (a_0, b_0) satisfying (1) does not exist. We then prove two theorems each giving a simple sufficient condition for the existence of such a pair. One or the other condition is almost always satisfied in practice. For example, the first requires, merely, that both of the sets $\{x \mid \frac{1}{3} \leqq F(x) \leqq \frac{2}{3}\}$ and $\{x \mid \frac{1}{3} \leqq G(x) \leqq \frac{2}{3}\}$ be nondegenerate. Next, we show that in any case if the set of minimizing pairs is nonempty then it is convex. This fact is used to obtain a fairly precise description of the set of minimizing pairs for the case F is increasing and continuous. In this case, simple conditions on G imply the uniqueness of a minimizing pair. Applications, especially to an estimation problem involving an unknown scale and location parameter, are then discussed.

Throughout the paper, the right hand side of (1) is denoted by M. Also, F and G are understood to be continuous on the right.

2. An example. Let F be the normal distribution function with mean 0 and variance 1. Let $6G(x) = 1 - (1 - e^{x})^{1/2}$, if $x < 0$, $= 5 + (1 - e^{-x})^{1/2}$, if $x \geqq 0$. Here, $M = \frac{1}{3}$ since $M \geqq \frac{1}{3}$ and

$$\sup |F(ax) - G(x)| = \sup_{0<x<1/a} [G(x) - F(ax)] \leqq G(1/a) - F(0)$$

which approaches $\frac{1}{3}$ as a increases indefinitely. For any pair (a, b),

$$\sup |F(ax + b) - G(x)| \geqq \sup |F(ax) - G(x)| = \sup [G(x) - F(ax)] > \tfrac{1}{3}$$

since $G(x) - F(ax)$ approaches $\frac{1}{3}$ as x approaches 0 through positive values of x and there is an $\epsilon > 0$ such that the derivative with respect to x of $G(x) -$

Received July 5, 1958; revised December 29, 1958.

$F(ax)$ is positive if $0 < x < \epsilon$. Thus a pair (a_0, b_0) satisfying (1) does not exist here.

3. Sufficient conditions for existence.

THEOREM 1. *If both of the sets* $\{x \mid \frac{1}{3} \leqq F(x) \leqq \frac{2}{3}\}$ *and* $\{x \mid \frac{1}{3} \leqq G(x) \leqq \frac{2}{3}\}$ *are nondegenerate then there is a real number pair* (a_0, b_0), *with* $a_0 > 0$, *satisfying* (1).

LEMMA. *If A is a closed bounded set of positive numbers and B is a closed bounded set of real numbers then there is a number a_0 in A and a number b_0 in B such that*

$$(2) \qquad \sup_{-\infty<x<\infty} | F(a_0 x + b_0) - G(x) | = \inf_{\substack{a\varepsilon A \\ b\varepsilon B}} \sup_{-\infty<x<\infty} | F(ax + b) - G(x) | .$$

PROOF OF LEMMA. Let M' denote the right hand side of (2). There is a sequence $a_0, a_1, \cdots$ in A and a sequence $b_0, b_1, \cdots$ in B such that

$$\sup |F(a_n x + b_n) - G(x)| < M' + 1/n$$

for each positive integer n and by the Bolzano-Weierstrass theorem these sequences can be chosen so that

$$(3) \qquad \lim_{n\to\infty} a_n = a_0, \qquad \lim_{n\to\infty} b_n = b_0 .$$

Let S be the set of numbers z such that $F(a_0 x + b_0)$ is continuous in x at z. If z is in S then $|F(a_n z + b_n) - G(z)| < M' + 1/n$ for each positive integer n and thus, by (3), $|F(a_0 z + b_0) - G(z)| \leqq M'$. Since S is a dense subset of the real numbers we have that $\sup |F(a_0 x + b_0) - G(x)| \leqq M'$ for x real which implies the desired result.

PROOF OF THEOREM 1. We first prove that $M \leqq \frac{1}{3}$. The assumptions imply the existence of numbers p_1, p_2, q_1, q_2 such that $p_1 < p_2$, $q_1 < q_2$, $F(p_i -) \leqq i/3 \leqq F(p_i)$, $G(q_i -) \leqq i/3 \leqq G(q_i)$, $i = 1, 2$. Thus, each of

$$F\left(\frac{p_2 - p_1}{q_2 - q_1} x + \frac{p_1 q_2 - p_2 q_1}{q_2 - q_1}\right)$$

and $G(x)$ is in the interval $[0, \frac{1}{3}]$ if $x < q_1$, each is in $[\frac{1}{3}, \frac{2}{3}]$ if $q_1 \leqq x < q_2$, and each is in $[\frac{2}{3}, 1]$ if $q_2 \leqq x$. This implies that $M \leqq \frac{1}{3}$.

If $M = \frac{1}{3}$ then the desired pair (a_0, b_0) exists by the above paragraph. Suppose $M < \frac{1}{3}$. Let $M < N < \frac{1}{3}$. There are numbers $r_0, \cdots, s_3$ such that $r_0 < r_1 < r_2 < r_3$, $s_0 < s_1 < s_2 < s_3$, $\frac{1}{3} \leqq F(r_i) \leqq \frac{2}{3}$, $\frac{1}{3} \leqq G(s_i) \leqq \frac{2}{3}$, $i = 1, 2$, and such that each of the numbers $F(r_1) - G(s_0)$, $G(s_1) - F(r_0)$, $F(r_3) - G(s_2)$, $G(s_3) - F(r_2)$ is greater than N. If (a, b) is a real number pair with $a > 0$ and at least one of the inequalities $r_0 < as_1 + b$, $r_3 > as_2 + b$, $s_0 < (r_1 - b)/a$, $s_3 > (r_2 - b)/a$ is not satisfied, or equivalently if

$$(4) \qquad \max\{r_0 - as_1, r_2 - as_3\} < b < \min\{r_3 - as_2, r_1 - as_0\}$$

is not satisfied, then

$$(5) \qquad \sup_{-\infty<x<\infty} |F(ax + b) - G(x)| > N.$$

For example, suppose $r_0 < as_1 + b$ is not satisfied. Then $r_0 \geqq as_1 + b$ and $|F(as_1 + b) - G(s_1)| \geqq G(s_1) - F(as_1 + b) \geqq G(s_1) - F(r_0) > N$, implying (5). Let $c_0 = (r_2 - r_1)/(s_3 - s_0)$ and $c_1 = (r_3 - r_0)/(s_2 - s_1)$. Then $0 < c_0 < c_1$. Let $A = [c_0, c_1]$, $d_0 = \inf_{a \varepsilon A} \max \{r_0 - as_1, r_2 - as_3\}$, $d_1 = \sup_{a \varepsilon A} \min \{r_3 - as_2, r_1 - as_0\}$, and $B = [d_0, d_1]$. If (a, b) is a real number pair with $a > 0$ such that either a is not in A or b is not in B, then (a, b) does not satisfy (4) and hence satisfies (5). For example, if $a > c_1$ then $r_0 - as_1 > r_3 - as_2$ and (4) is not satisfied. Therefore, $M = \inf_{a \varepsilon A, b \varepsilon B} \sup_{-\infty < x < \infty} |F(ax + b) - G(x)|$ and the desired result follows from the lemma.

THEOREM 2. *If F is continuous and G is a step function with at most a finite number of discontinuity points in each bounded interval, then there is a real number pair (a_0, b_0), with $a_0 > 0$, satisfying* (1).

PROOF. The only case that needs to be considered here is the one for which there is a number s satisfying $G(s-) < \frac{1}{3}$ and $\frac{2}{3} < G(s)$. The other case is taken care of by Theorem 1. Let $2K = G(s) - G(s-)$. The assumptions imply that $K \leqq M \leqq \max \{K, G(s-), 1 - G(s)\}$ and the desired pair (a_0, b_0) is easily seen to exist if M is equal to the right hand side of this expression. Suppose that M is less. Then there is a number N satisfying $M < N < \max \{G(s-), 1 - G(s)\} < \frac{1}{3}$. There are numbers $r_0, \cdots, s_3$ such that $r_0 < r_1 < r_2 < r_3$, $s_0 < s_1 < s_2 < s_3$, $s_1 < s < s_2$, $F(r_1) = \frac{1}{3}$, $F(r_2) = \frac{2}{3}$, $G(s_1) = G(s-)$, $G(s_2) = G(s)$ and such that each of the numbers $F(r_1) - G(s_0)$, $G(s_3) - F(r_2)$, and $\max \{G(s-) - F(r_0), F(r_3) - G(s)\}$ is greater than N. It follows that if a pair (a, b) with $a > 0$ does not satisfy both $as_0 + b < r_1 < as + b < r_2 < as_3 + b$ and $\max \{r_3 - (as_2 + b), as_1 + b - r_0\} > 0$, then $\sup |F(ax + b) - G(x)| > N$. It follows that there is a closed bounded positive number set A and a closed bounded real number set B such that

$$M = \inf_{a \varepsilon A, b \varepsilon B} \sup_{-\infty < x < \infty} |F(ax + b) - G(x)|$$

and the desired result follows from the lemma.

4. Convexity of the set of minimizing pairs.

Let Q be the set of minimizing pairs. That is, let Q be the set of all pairs (a_0, b_0), with $a_0 > 0$, satisfying (1).

THEOREM 3. *If Q is nonempty, then Q is convex.*

PROOF. Suppose that (c, d) and (e, f) are in Q and that $0 < \lambda < 1$. We are to show that (a, b) is in Q where $a = (1 - \lambda)c + \lambda e$ and $b = (1 - \lambda) d + \lambda f$. If x satisfies $cx + d \leqq ex + f$ then $cx + d \leqq ax + b \leqq ex + f$, $-M \leqq F(cx + d) - G(x) \leqq F(ax + b) - G(x) \leqq F(ex + f) - G(x) \leqq M$, and

$$|F(ax + b) - G(x)| \leqq M.$$

Similarly, the last inequality holds if x satisfies $cx + d > ex + f$. Thus, (a, b) is in Q.

THEOREM 4. *If F is increasing then there is a number t and a number k such that $at + b = k$ for each (a, b) in Q. If, in addition, F is continuous and Q contains more than one element then*

$$(6) \qquad |F(ax + b) - G(x)| < (G(t) - G(t-))/2 = M$$

for each $x \neq t$ and each (a, b) belonging to the interior of Q relative to the line that contains Q.

Thus, if F is increasing then Q is a convex subset of a nonvertical line. If F is increasing and continuous and Q contains more than one element then t is the unique number x maximizing $G(x) - G(x-)$, k is the unique number satisfying $2F(k) = G(t-) + G(t)$, M is easily calculated, and so forth.

PROOF. Suppose the first assertion is not true. Then either there are three points in Q not colinear or Q is a nondegenerate subset of a vertical line. Both possibilities imply, the former by the convexity of Q, that there are pairs (a, c) and (a, d) in Q such that $c < d$. Let $b = (c + d)/2$. Then (a, b) is in Q. Either $\sup [F(ax + b) - G(x)] = M$ or $\inf [F(ax + b) - G(x)] = -M$. Suppose the former is true, the other case being similar. Then there is an x_0 such that

$$\sup [F(ax + b) - G(x)] = \max \{F(ax_0 + b) - G(x_0),$$
$$F([ax_0 + b]-) - G(x_0-)\}.$$

Since F is increasing, the right hand side is less than the corresponding expression with b replaced by d which in turn is less than or equal to M. Thus, $\sup [F(ax + b) - G(x)] < M$, a contradiction. The first assertion follows.

Suppose that F is continuous as well as increasing. Then, if (a, b) is in Q,

$$(7) \quad \inf_{-\infty < x < \infty} [F(ax + b) - G(x)] = -M, \qquad \sup_{-\infty < x < \infty} [F(ax + b) - G(x)] = M.$$

For suppose otherwise. For example, suppose the second relation is not true. Then $\sup [F(ax + b) - G(x)] < M$ and necessarily the first relation is true. The assumptions on F imply that the left hand side of each relation is continuous and increasing in b. Hence, there is a number $c > b$ such that $\inf [F(ax + c) - G(x)] > -M$ and $\sup [F(ax + c) - G(x)] < M$. Thus, $\sup |F(ax + c) - G(x)| < M$, a contradiction.

Now suppose that Q contains more than one point and that (a, b) is a relative interior point of Q. Then, if $\epsilon > 0$,

$$(8) \quad \begin{aligned} -M &< \inf_{|x-t|>\epsilon} [F(ax + b) - G(x)], \\ &\sup_{|x-t|>\epsilon} [F(ax + b) - G(x)] < M. \end{aligned}$$

Suppose, for example, that the second inequality is not true. Then there is an $x_0 \neq t$ such that $F(ax_0 + b) - G(x_0-) = M$. By assumption there are pairs (a_1, b_1) and (a_2, b_2) in Q such that $a = (a_1 + a_2)/2$ and $b = (b_1 + b_2)/2$. Since $x_0 \neq t$, $ax_0 + b$ is strictly between $a_1x_0 + b_1$ and $a_2x_0 + b_2$. Therefore, $M \geqq \max_{i=1,2} [F(a_ix_0 + b_i) - G(x_0-)] > F(ax_0 + b) - G(x_0-)$, a contradiction. By (7) and (8), $M = F(k) - G(t-) = G(t) - F(k)$. Therefore, $2M = G(t) - G(t-)$ and the desired inequality in (6) follows from (8).

COROLLARY. *Suppose that F is increasing and continuous and that G is either* (i) *continuous, or* (ii) *a step function with $n > 1$ discontinuity points at which G*

has jumps of size $1/n$. *Then there is a unique pair* (a_0, b_0), *with* $a_0 > 0$, *satisfying* (1).

Remark. The condition on G is more special than need be. It could be replaced by the following condition: *As x varies $G(x) - G(x-)$ assumes its maximum value at least twice.*

The corollary and remark are immediate consequences of Theorems 1 and 4.

5. Discussion. In practically all cases of interest the results of section 3 imply that if G is a distribution function then there is, for example, a best normal approximation to G.

We now discuss an estimation problem invoving a scale and location parameter. Suppose that F is an increasing and continuous distribution function. Let $n > 1$ and suppose that $X_1, \cdots, X_n$ are independent random variables each with the distribution function $F(\cdot; \mu, \sigma)$ where $F(x; \mu, \sigma) = F([x - \mu]/\sigma)$ for all real x. The parameter (μ, σ) is an element of the parameter space Ω which is here taken to be the open upper half plane. Since F is continuous we may restrict ourselves to the set $\mathfrak{X}$ of sample points $(x_1, \cdots, x_n)$ with all coordinates distinct. For each $(x_1, \cdots, x_n)$ in $\mathfrak{X}$ let $G(\cdot; x_1, \cdots, x_n)$ be the corresponding empirical distribution function. We ask whether or not there is an estimate $\delta = (\delta_1, \delta_2)$ of (μ, σ) such that

$$\begin{aligned}(9)\quad &\sup_{-\infty<x<\infty} |F(x; \delta_1(x_1, \cdots, x_n), \delta_2(x_1, \cdots, x_n)) - G(x; x_1, \cdots, x_n)| \\ &\qquad = \inf_{(\mu,\sigma)\in\Omega} \sup_{-\infty<x<\infty} |F(x; \mu, \sigma) - G(x; x_1, \cdots, x_n)|\end{aligned}$$

for each $(x_1, \cdots, x_n)$ in $\mathfrak{X}$. Such an estimate would be a minimum distance estimate in the terminology of Wolfowitz who has studied the role of the empirical distribution function in estimation in very general contexts (See [1] and also the references listed in [1]). The corollary of the previous section implies that a function δ on $\mathfrak{X}$ satisfying (9) does exist and is unique. One question arises. Is δ measurable? It is easy to prove even more: δ is continuous.

REFERENCE

[1] J. Wolfowitz, "The minimum distance method," *Ann. Math. Stat.*, Vol. 28 (1957), pp. 75–88.