SEQUENTIALLY LOCALIZABLE FUNCTIONALS

by

RAYMOND S. BOOTH
B.Sc. University of Otago, New Zealand, 1961
M.Sc. University of Otago, New Zealand, 1962

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

in the Department
of
MATHEMATICS

We accept this thesis as conforming to the
required standard

University of British Columbia
December, 1965

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Mathematics.

The University of British Columbia
Vancouver 8, Canada

Date 20th. December 1965

The University of British Columbia

FACULTY OF GRADUATE STUDIES

PROGRAMME OF THE

FINAL ORAL EXAMINATION

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

of

RAYMOND SIDNEY BOOTH

B.A., University of Otago, 1961

M.Sc., University of Otago, 1962

THURSDAY, DECEMBER 16th, 1965, AT 3:30 P.M.

IN ROOM 225, MATHEMATICS BUILDING

COMMITTEE IN CHARGE

Chairman: I. McT. Cowan

P. Bullen	I. Hacking
C. Froese	Z. A. Melzak
E. E. Granirer	B. N. Moyls

Research Supervisor: Z. A. Melzak

External Examiner: J. E. L. Peck

University of Alberta,

Calgary

SEQUENTIALLY LOCALIZABLE FUNCTIONALS

ABSTRACT

A standard way of finding the unique zero on $(0, 1)$ of a continuous decreasing function, with $f(0)\, f(1) < 0$, is to test the sign of $f(1/2)$, then the sign of $f(1/4)$ (if $f(1/2) < 0$) or the sign of $f(3/4)$ (if $f(1/2) > 0$), etc. In this way, the zero of f is localized in n steps to an interval of length 2^{-n}. The unique maximum of a unimodal function on $[0, 1]$ can be similarly localized, but the unique maximum of a unimodal function on the unit square cannot. We start by generalizing these problems: let A be a compact subset of E^n, let F be a set of real-valued functions on A, and for each f in F let $S(f)$ be a point in the Cartesian product A^k; $S(f)$ is called a functional on F. Examples of such functionals are zeros, extrema, inflexion points, saddle points, etc., as well as sets of these. A test function T is a function of m real variables, which takes up only a finite number (≥ 2) of distinct values. An abscissa set X_i is an ordered m-tuple $(x_{1i}, \ldots, x_{mi})$, with each x_{ji} in A. A sequential strategy is a way of selecting abscissa sets $X_1, X_2, \ldots,$

where the knowledge of $T(f(x_{1i}),\ldots,f(x_{mi}))$ is used to determine X_{i+1}. The N-th set of indeterminacy for S(f) is the largest subset of A^k in which S(f) can lie, consistent with the results of the first N-1 tests. A functional S(f) is sequentially localizable if a test function T and a sequential strategy exist, such that for every f in F the sets of indeterminacy shrink to a point (which must then be S(f) itself).

First, several conditions are given to ensure the sequential localizability of a functional, these are presented in terms of certain topologies induced on A^k and in terms of contraction maps. It is then shown that if a functional is localizable, there exists an optimal strategy under which the sets of indeterminacy converge fastest; further, the speed of localization is always exponential. Next, the concept of a random strategy and of random localizability is introduced, and it is shown that in many cases random localizability and sequential localizability are equivalent. Also, the speed of the former is not too much worse than the speed of the latter. Finally, optimal and near-optimal strategies are worked out for some functionals of interest.

GRADUATE STUDIES

Differential Equations	C. Clark
Functional Analysis	D. Bures
Complex Variable	R. Cleveland
Convexity	Z. A. Melzak
Differential Geometry	E. Luft

ABSTRACT

Supervisor: Dr. Z.A. Melzak

A standard way of finding the unique zero of a continuous decreasing function f(x) defined on the closed interval [0,1], where $f(0) > 0 > f(1)$, is the bisection method. One tests the sign of $f(\frac{1}{2})$; if this is positive, then one tests the sign of $f(\frac{3}{4})$, and if negative, one tests $f(\frac{1}{4})$, etc. In this way, the zero is localized in n steps to an interval of length 2^{-n}. The unique maximum point of a unimodal function defined on [0,1] can be localized in a similar manner; however, the maximum point of a unimodal function defined on a square cannot be so localized.

We generalize these problems. Let D be a compact subset of Euclidean q-dimensional space, and let F be a family of real-valued functions defined on D. To each $f \in F$, associate a unique point S(f) in $D \times D \times D \times \ldots \times D$ (m times).

The mapping S is called a functional. An abscissa set is an ordered k-tuple $X = (x_1, x_2, \ldots, x_k)$ where $x_i \in D$. A strategy is a rule for selecting a sequence of abscissa sets $X_0, X_1, X_2, \ldots, X_n, \ldots$, where each X_n depends on X_i, and the values of a function $f \in F$ at the elements of X_i, $i = 0, 1, \ldots (n-1)$. The functional is said to be localizable if there is a strategy such that the diameter of the largest set K_n which may contain S(f), consistent with the observations of the values of f at the abscissa sets X_i, $i = 0, 1, \ldots, n-1$, decreases to zero as $n \to \infty$, independently

of the function $f \in F$.

A test function is a map T, with finite range V, which associates with each abscissa set X and the values of a function f at the elements of X, an element of V. It is shown in chapter 2 that many functionals can be localized with the aid of a suitable test function. Several necessary and sufficient conditions for a functional to be localizable are given. In particular, localization is characterized in terms of the strength of a certain topology induced on D^m by the strategy or test function.

In chapter 3, the speed of localization is considered, and it is shown that in most cases the speed is exponential. A measure of the speed is defined, and sufficient conditions are given for the existence of an optimal strategy.

Chapter 4 is concerned with random strategies and random localizability, and it is shown that localizability often implies random localizability. The particular problem of locating the zero of a continuous monotone function defined on an interval by random localization is considered in detail, and compared with the classic bisection method.

Finally, optimal and near-optimal strategies are worked out for a number of functionals of special interest.

TABLE OF CONTENTS

		page
CHAPTER 1	INTRODUCTION	1
1 - 1	The bisection problem	1
1 - 2	Outline of the general problem	1
1 - 3	The unimodal problem	7
1 - 4	An example of non-localizability	11
1 - 5	A two-dimensional problem	13
1 - 6	Remarks	22
CHAPTER 2	LOCALIZABILITY	25
2 - 1	Basic definitions	25
2 - 2	Test functions	27
2 - 3	Examples	31
2 - 4	L-topology	34
2 - 5	A simultaneous search problem	38
CHAPTER 3	RATE OF LOCALIZATION	45
3 - 1	Speed	45
3 - 2	Speed with Lipschitz Conditions	53
3 - 3	Exponents	56
3 - 4	Optimal Strategies	63

		page
CHAPTER 4	RANDOM LOCALIZABILITY	68
4 - 1	Definitions	68
4 - 2	General theorems and examples	72
4 - 3	Random search for a zero	76
4 - 4	Joint probability distribution	79
4 - 5	Expected values	91
4 - 6	The behaviour of $e_n(\alpha)$	99
4 - 7	Rectangular density function	118
CHAPTER 5	LOCALIZATION OF ZEROES OF DERIVATIVES	121
5 - 1	General results	121
5 - 2	Derivatives of odd order	127
BIBLIOGRAPHY		130

ACKNOWLEDGEMENTS

I wish to thank my supervisor, Dr. Z.A. Melzak, for suggesting the topic for this thesis, and for providing much valuable guidance throughout my graduate studies and during the preparation and writing of this thesis. The financial assistance of the Canadian Universities' Foundation via a scholarship which made it possible for me to study in Canada is gratefully acknowledged.

CHAPTER 1

INTRODUCTION

1 - 1 The bisection problem

One of the oldest methods of approximating the zero of a function is the bisection method. Suppose f is any continuous real-valued function defined on the closed interval $[0,1]$ of the real line such that $f(0) > 0 > f(1)$, f possessing a unique zero α in this interval. By computing the sign of $f(\frac{1}{2})$, it can be determined whether $0 < \alpha < \frac{1}{2}$, $\alpha = \frac{1}{2}$, or $\frac{1}{2} < \alpha < 1$; the first of these holds if $f(\frac{1}{2}) < 0$, the second if $f(\frac{1}{2}) = 0$, and the third if $f(\frac{1}{2}) > 0$. Ignoring the possibility that $f(\frac{1}{2}) = 0$, when clearly no further computation is needed, the possible position of α is restricted to an interval of length at most $\frac{1}{2}$.

By considering the sign of f at the mid-point of the new interval, one can in the same way, restrict the possible position of α to an interval of length at most $\frac{1}{4}$. By successively computing the sign of f at the mid-point of the largest interval to which α can belong, on the basis of previous evaluations of f; one can, after n steps involving the evaluation of f at n points, restrict α to an interval of length at most 2^{-n}. Thus by taking n sufficiently large, α can be approximated as accurately as desired.

1 - 2 Outline of the general problem

More generally, let F be a family of real-valued functions defined on a given domain D. Suppose that the values of

$f(x)$, $x \in D$, $f \in F$, are all unknown, but can always be computed for any given $x \in D$. It is desired to locate a particular point associated with the function f (e.g., a zero, a maximum-point, a point of inflexion) as accurately as possible by finding the values of $f(x)$ for a suitable sequence of properly selected points $x \in D$. The mapping which assigns to each member f of F a point, will be called a functional; the image of the mapping will be written $S(f)$. $S(f)$ can be regarded as a point in the Cartesian product $D \times D \times D \ldots \times D$ (m times), where m is a positive integer. This would allow consideration of the problem of approximating m zeros of a suitable function simultaneously.

The problem of finding methods for approximating such a point has been considered for special cases since before the time of Newton, especially if severe restrictions are placed on F. For example, there are a very large number of ways in which one may approximate a solution to the equation $f(x) = 0$ if f is a polynomial in x of small degree.

We are concerned here with problems which can be solved by generalizing the bisection method, and, with this in mind, make the following general assumtpions.

(i) The domain D is a compact convex subset of finite-dimensional Euclidean space. The class of functions F is usually taken to be the largest possible class for which the problem is well-defined, provided non-trivial results may be obtained. The class F for the bisection problem, i.e., the class of all real continuous functions f defined on $D = [0,1]$ such that $f(0) > 0 > f(1)$, possessing a unique zero α, $0 < \alpha < 1$; is typical of the type of class F to be considered.

(ii) If $x \in D$ is given, and $f \in F$, then no errors are made in

computing the numerical value of $f(x)$. This property means that some care is needed in practice when using the subsequent results to be obtained. The case of errors can be handled in some instances by stochastic approximation methods as shown by Robbins and Monro [8], but we shall not be concerned with this subject here.

(iii) The choice of points x in D for which $f(x)$ is computed is to be determined on a sequential basis. That is, suppose n is a positive integer, and already n points $x_1, x_2, \ldots x_n \in D$ have been selected, and $f(x_1), f(x_2), \ldots f(x_n)$ have been evaluated, for some $f \in F$. The choice of the next point $x_{n+1} \in D$ may depend on the points $x_1, x_2, \ldots x_n$, and the numerical values $f(x_1), f(x_2), \ldots f(x_n)$. For example, in the bisection problem, we choose $x_1 = \frac{1}{2}$ and evaluate $f(\frac{1}{2})$. We then choose $x_2 = \frac{1}{4}$ if $f(\frac{1}{2}) < 0$; $x_2 = \frac{3}{4}$ if $f(\frac{1}{2}) > 0$.

(iv) A rule for determining the selection of the successive elements x_n in D at which a function $f \in F$ is to be evaluated shall be given before any computations are made. Such a rule will be called a strategy. The strategy for the bisection problem is simple; find the largest interval to which the zero of the function f can belong on the basis of previous evaluations (if any), and select the mid-point of the interval. There are, of course, many alternative strategies which could be used to approximate the zero of f. One of the points of trisection would suffice.

(v) Before any computations are made, it must be known that all points $S(f)$, $f \in F$, belong to some set K_o of (bounded) diameter d_o. K_o is a subset of a Cartesian product of D with itself a finite number of times. The diameter d_o is that defined by the usual Euclidean metric. For the bisection problem we may take K_o as the interval D, with $d_o = 1$.

(vi) For a given strategy M, a function $f \in F$ and any positive integer n, we can define a set $K_n(f,M)$ which is known to contain S(f), after f has been evaluated at n sets of points in D determined successively by the strategy M. $K_n(f,M)$ is a (possibly improper) subset of K_o. As a measure of the worth of the strategy, and of the speed of the determination of the functional, we set

$$d_n'(M) = \sup_{f \in F} d[K_n(f,M)]$$

where $d[K_n(f,M)]$ is the diameter of the set $K_n(f,M)$. We shall say that the functional is localizable with respect to the strategy M if

$$\lim_{n \to \infty} d_n'(M) = 0 ;$$

this means that S(f) can be successively restricted to sets $K_n(f,M)$ whose diameters decrease to zero. For example, in the bisection problem, we have

$$d_n'(M) \leq 2^{-n} \quad \text{for all positive integers n.}$$

So, for this example, the functional is localizable by the bisection strategy.

It is sometimes convenient to say that S(f) is localizable. No confusion arises, since f is a typical member of F, and the set of all points S(f) for $f \in F$, is just the range of the functional.

We can now raise some questions of a general nature concerning localizability.

(i) For given D and F, is there a strategy M with respect to which the functional is localizable ?

(ii) If there does not exist such an M, can F be restricted

to a reasonably large subclass F' of F, where localization is now possible?

(iii) If localization is possible, can one determine a best strategy in some sense? If, by best we mean one for which the diameters $d_n'(M)$ decrease fastest, then for the bisection problem, the best strategy is easily seen to be the one described in section 1-1.

It is possible to take other factors into consideration in deciding what is to be regarded as optimal. For example, the cost of evaluating $f(x)$ for some $f \in F$, $x \in D$, computer time involved, or the calculations used to determine the successive points at which f is to be evaluated, could all be taken into account. Even if we ask for a minimal $d_n'(M)$, there are two distinct cases to be considered:

(i) strategies for which an a priori upper bound N is given on the number of points in D at which a function may be evaluated,

(ii) strategies which are allowed to continue indefinitely.

Should we ask for a strategy M (if there is one) which minimizes $d_N'(M)$, where N is decided in advance of any computations; the strategy M may not be so effective if, after evaluating a function f at N points in D, we then decide to localize more exactly by further computation. Kiefer,[6], defines optimality for the case where N is given in advance, using the term minimax, or ϵ-minimax, and considers optimal strategies for some of the problems we discuss later. His treatment of the problem however, differs from ours. In general, we are concerned with strategies for which the number of observations is allowed to increase indefinitely.

We can also ask for optimal strategies under the condition that computations cease as soon as the point being localized

is known to belong to a set with diameter at most ϵ, where ϵ is given.

A strategy of a different type will concern us in chapter 4. Rather than selecting the points $x_1, x_2, \ldots, x_n$ in D by a fixed rule, we consider the possibility of selection 'by chance'. As an illustration, we take the bisection problem, with $D = [0,1]$, and define a density function $\phi(t)$, $0 \leq t \leq 1$. Instead of selecting $x_1 = \frac{1}{2}$, and evaluating $f(\frac{1}{2})$, where $f \in F$, suppose we select x_1 arbitrarily in D, in such a way that for all real numbers a,b with $0 \leq a < b \leq 1$, the probability that $a < x_1 < b$ is

$$\int_a^b \phi(t)dt.$$

As before, we have

$$x_1 < S(f) < 1 \quad \text{if } f(x_1) > 0$$
$$0 < S(f) < x_1 \quad \text{if } f(x_1) < 0 .$$

So, S(f) can be restricted to one of the two intervals $[0,x_1]$ or $[x_1,1]$. With the density function redefined on the new interval, we can similarly select a point x_2, and continue this procedure indefinitely. We then ask, what is the expected value of the length of interval to which S(f) is confined after f is evaluated at n points in D? One of our aims is to compare this with the bisection strategy. R.S. Pinkham [7] has considered this problem with the restriction that one stops computation as soon as S(f) is localized to an interval of length h, where h is given.

We postpone definitions and theorems until chapters 2 and 3. The remainder of this chapter gives some examples to illustrate the previous remarks.

1 - 3 Unimodal problem

For completeness and future reference, we insert the problem of localizing the maximum (or minimum) point of a unimodal function. This was first solved by Kiefer [5] in 1953, and is dealt with in detail by Bellman and Dreyfus [1].

Let $D = [a,b]$ be a compact interval on the real line. Let F be the class of all continuous unimodal real-valued functions defined on D which possess a (unique) maximum point. That is, if $f \in F$, then

$$f \text{ is strictly increasing on } [a, S(f)]$$
$$f \text{ is strictly decreasing on } [S(f), b]$$

for some point $S(f)$ in D. We allow the improper case $S(f) = a$ or $S(f) = b$. The problem is to localize the maximum point at $S(f)$.

The strategy uses the fact that if $f \in F$, and x,y are any two points in D, with $x < y$, then

$$a \leq S(f) < y \quad \text{if} \quad f(x) > f(y)$$
$$x < S(f) \leq b \quad \text{if} \quad f(x) < f(y)$$
$$x < S(f) < y \quad \text{if} \quad f(x) = f(y).$$

If these conditions do not hold, then there would be more than one maximum point, contrary to hypothesis.

Thus we can always localize $S(f)$ to a subinterval of D of length $\leq \text{Max } \{y-a, b-x\}$, and this is a proper subinterval of D of length $\leq c(b-a)$, where c is a fixed constant less than 1, independent of a and b, provided x and y are properly chosen. By iteration, we can successively restrict $S(f)$ to a sequence of intervals whose lengths decrease to zero. So the functional is localizable.

In order to find the best strategy, that is, the strategy M which minimizes $d_N(M)$, N being the number of points in D at

which f may be evaluated; it is convenient to ask the following question. If D is the interval $[0, L_N]$, what is the largest possible value of L_N, so that S(f) can be restricted to a unit interval by computing f at most N points in D ? The answer is given by the following theorem.

Theorem 1.1 If the sequence F_n is defined by the equations

$$F_o = F_1 = 1$$

$$F_n = F_{n-1} + F_{n-2} \quad \text{for all integers } n \geq 2,$$

then $L_n \leq F_n$ for all positive integers n. Moreover, if $\epsilon > 0$ is arbitrary, there exists a strategy with respect to which S(f) can be localized to an interval of unit length, if

$$L_n = \frac{F_n}{1+\epsilon} \qquad n \geq 2$$

$$L_o = L_1 = 1 .$$

F_n is the nth Fibonaccian number.

We omit the proof of this theorem since an excellent account of it is given by Bellman and Dreyfus [1]. However, we wish to give an outline of the proof since the proofs of some later theorems are based on similar arguments.

That $L_o = L_1 = 1$ is obvious, because evaluating f at just one point in D yields no information about S(f). Suppose n is an integer ≥ 2, and two points x_1, x_2 are selected so that

$$0 \leq x_2 < x_1 \leq L_n .$$

One can by evaluating f at x_1 and x_2, restrict S(f) to either $[0, x_1]$ or $[x_2, L_n]$. Since, to restrict S(f) further, a third point x_3 must be chosen, the inequalities

$$x_1 \leq L_{n-1} \tag{1.1}$$

$$L_n - x_2 \leq L_{n-1} , \tag{1.2}$$

must hold if S(f) can be localized to a unit interval, by

evaluating f at at most n points in D. No matter where a third point is chosen, the inequalities

(1.3) $$x_2 \leq L_{n-2}$$

(1.4) $$L_n - x_1 \leq L_{n-2}$$

must also hold, otherwise again more than n points would be required. So, by addition of (1.2) and (1.3)

(1.5) $$L_n \leq L_{n-1} + x_2$$

$$L_n \leq L_{n-1} + L_{n-2}$$

for all integers $n \geq 2$, which gives

$$L_n \leq F_n$$

for all non-negative integers n.

The second part of the theorem follows from the fact that if f is defined on $[0,L_n]$, $n \geq 2$, we can by choosing $x_1 = L_{n-1}$ and $x_2 = L_{n-2}$ and evaluating f at these points, restrict S(f) to an interval of length L_{n-1}. We need choose only one new point x_3, to restrict S(f) to an interval of length L_{n-2}, since the points x_2, x_1 are interior respectively to $[0,x_1]$ or $[x_2,L_n]$. By iteration, we can repeat until S(f) is restricted to an interval of length 2, where the value of f is known at the mid-point of this interval, and one final selection x_n may be made. Since we cannot choose x_n at the mid-point, we choose x_n at an arbitrary distance ϵ from the mid-point, allowing one to restrict S(f) to an interval of length $1+\epsilon$.

By a change of scale by the factor $1+\epsilon$, the result follows.

Corollary 1.2 If F is initially defined on the interval $D = [0,1]$, if $\epsilon > 0$ and an integer $n \geq 2$ are given, then for any $f \in F$, S(f) can be localized to an interval of length $(1+\epsilon)/F_n$ by

evaluating f at a sequence of n suitable points in D.

Remark 1.3 The integer n must be given a priori, since the selection of the first two points x_1 and x_2 depend on n. Indeed, if D = [0,1],

$$x_1 = F_{n-1}/F_n \quad \text{and} \quad x_2 = F_{n-2}/F_n .$$

However, another strategy for localizing S(f) suggests itself, which does not require that n be decided in advance. It is well known that

$$F_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^{n+1} \qquad n \geq 0,$$

and so

(i) $$F_n \sim \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{n+1}$$

(ii) $$\lim_{n \to \infty} F_n/F_{n-1} = \frac{1+\sqrt{5}}{2} \approx 1.618 .$$

This leads to what we will refer to as the golden-mean strategy. Starting with domain D = [0,1], we select the two points

$$x_1 = \frac{\sqrt{5}-1}{2} \qquad x_2 = 1 - \frac{\sqrt{5}-1}{2}$$

which divide D in the golden ratio, and for some $f \in F$, evaluate $f(x_1)$ and $f(x_2)$. We than have either

$$S(f) \in [0,x_1] \text{ with } f(x_2) \text{ known}$$
$$\text{or } S(f) \in [x_2,1] \text{ with } f(x_1) \text{ known.}$$

Both these intervals have length $\frac{\sqrt{5}-1}{2}$, and in either case, one new point x_3 can be chosen so that again we have two points which divide the appropriate interval in the golden ratio. By iterating this indefinitely, we can restrict S(f) to an interval

of length at most $\left(\frac{\sqrt{5}-1}{2}\right)^{n-1}$, evaluating f at $n \geq 2$ suitable points in D.

1 - 4 An example of non-localizability

We show that if D and F are given, F being a union of two classes F' and F", then the functional defined by F is not necessarily localizable, even if localization is possible when F is intersected with either F' or F". In essence, this can happen because for every strategy which one might use to localize the functional, there exist functions $f \in F$ for which there is no way of deciding whether $f \in F'$ or $f \in F"$.

Let D be the closed unit interval [0,1]. Define F' to be the class of all continuous real-valued unimodal functions f with domain D, which possess a unique maximum at a point S(f), $0 < S(f) < 1$. Define F" to be the class of all continuous real-valued unimodal functions f with domain D, which possess a unique minimum at a point S(f), $0 < S(f) < 1$.

Define F to be the union of F' and F". Each member of F has a unique extreme point in the interior of D. In contrast to the functions of theorem 1.1, we do not allow the cases $S(f)=0$ or $S(f)= 1$; hence the two classes F' and F" are disjoint.

Now, by the results of section 1-3, if $f \in F$, we can localize S(f) by the golden-mean strategy if it is known that $f \in F'$. We can also localize S(f) if it is known that $f \in F"$. Thus to show that the functional is not localizable, we must demonstrate that, for every strategy, there is a function $f \in F$ for which there is insufficient information to decide whether $f \in F'$ or $f \in F"$. This is a consequence of the following lemma.

Lemma 1.4 Let $\epsilon > 0$ be given, and let k be an arbitrary positive integer. For any set of k distinct points $x_1, x_2, x_3, \ldots x_k$ in D such that

$$0 \leq x_1 < x_2 < x_3 < \ldots < x_k \leq 1$$

there exist two functions f and g satisfying

$$\begin{aligned} &(i) \quad f \in F' \qquad g \in F'' \\ &(ii) \quad f(x_i) = f(x_i) \qquad i = 1, 2, \ldots k \\ &(iii) \quad |S(f) - S(g)| \geq 1-\epsilon. \end{aligned}$$

Proof Let

$$\begin{aligned} u &= \min \{x_1, \epsilon\} \quad \text{if } x_1 > 0 \\ u &= \min \{x_2, \epsilon\} \quad \text{if } x_1 = 0 \\ v &= \max \{x_k, 1-\epsilon\} \quad \text{if } x_k < 1 \\ v &= \max \{x_{k-1}, 1-\epsilon\} \quad \text{if } x_k = 1 \end{aligned}$$

Define f and g by the equations

$$f(x) = \begin{cases} x & \text{if} \quad 0 \leq x \leq v \\ 4x - 3v & \text{if} \quad v \leq x \leq \frac{1+v}{2} \\ 3 - 2x & \text{if} \quad \frac{1+v}{2} \leq x \leq 1 \end{cases}$$

$$g(x) = \begin{cases} -2x & \text{if} \quad 0 \leq x \leq \frac{u}{2} \\ 4x - 3u & \text{if} \quad \frac{u}{2} \leq x \leq u \\ x & \text{if} \quad u \leq x \leq 1 \end{cases}$$

It is easy to check that f and g satisfy the requirements of the lemma.

Corollary 1.5 The functional defined by the set of all extreme points of the members of F is not localizable. For, since ϵ and k are arbitrary, we cannot guarantee to reduce the diameter of the smallest set which may contain S(f) to any number less than 1.

If we exclude the possibility that for $f \in F$, S(f) may be arbitrarily near either end of D, then localization is possible.

For it is then easy to determine whether $f\in F'$ or $f\in F''$ by evaluating f at two points of a subinterval of D which contains an end-point of D, but cannot contain S(f). This argument gives the following lemma whose proof we omit.

Lemma 1.6 Let δ, $0 < \delta < 1$ be given. Let F_δ be the subclass of F for which $0 < \delta \leq S(f)$ for all $f\in F_\delta$. Then the functional determined by the extreme points of members of F_δ is localizable.

If such a δ is not known, we can always attempt to localize S(f), under the two separate assumptions $f\in F'$ and $f\in F''$, by the golden-mean strategy. After evaluating f at (n+1) points of D under each assumption, we get two closed intervals $K_n'(f)$, $K_n''(f)$ in D, each of length $\leq \left(\frac{\sqrt{5}-1}{2}\right)^n$, such that

$$S(f) \in K_n'(f) \quad \text{if} \quad f\in F'$$
$$S(f) \in K_n''(f) \quad \text{if} \quad f\in F'' .$$

If one of these intervals does not contain an end-point of D, we can apply lemma 1.6, and determine whether $f\in F'$ or $f\in F''$. If not, the most that we can say is that

$$S(f) \in K_n'(f) \cup K_n''(f),$$

where $K_n'(f)$ contains one end-point of D, and $K_n''(f)$ contains the other.

1 - 5 A two-dimensional problem

We wish now to consider the problem of localizing the maximum points of a suitable class of functions defined on a square in the Euclidean plane. New difficulties arise in problems where the dimension of the domain D is greater than 1, not all of which are due to the lack of ordering of points in D. Generally speaking, the obvious analogues of the one-dimensional case do not put

sufficient restrictions on the class F in order to get localizability. We take three different classes of functions for the present problem, to show what situations may arise under the different assumptions.

Let D be the unit square

$$D = \{(x,y) \mid 0 \leq x \leq 1, \quad 0 \leq y \leq 1\}$$

in the Euclidean plane. The three classes of functions to be considered are defined as follows:

(i) F_1 is the class of all real continuous functions defined on D which possess a unique (relative) maximum point in the interior of D.

(ii) F_2 is the subclass of F_1 such that for every real number β, and all $f \in F_2$, the set $\{x \in D \mid f(x) \geq \beta\}$ is a strictly convex figure. Alternatively, we may say that for every real number β, the level curve $\{x \in D \mid f(x) = \beta\}$ is a convex curve containing no line segments.

(iii) F_3 is the subclass of F_1 such that for some fixed constant $\sigma > 0$, every real number β, and all $f \in F_3$, $\{x \in D \mid f(x) > \beta\}$ has diameter d, implies that $\{x \in D \mid f(x) > \beta\}$ contains a disc of radius σd. This requirement can be rephrased by the statement: all level curves are of 'bounded eccentricity'.

It is understood that the terms distance, diameter, etc., are referred to the usual Euclidean metric. In all three cases, S(f) is to be the point in D where f attains its maximum value.

The conditions on F_1 are quite weak, and it is almost obvious that localization is not possible. Indeed, we show below that if $f \in F_1$, then no matter how many points in D are chosen for evaluating f, S(f) may still be anywhere in D except on a finite

number of line segments. The conditions on F_2 are the direct analogue of those in theorem 1.1. The assumption that the level curves do not contain line segments is simply to ensure that if three points x_1, x_2, x_3 in D lie on a line, with x_2 between x_1 and x_3, and if $f(x_1) = f(x_3)$, then $f(x_2) > f(x_1) = f(x_3)$. We shall show that if $f \in F_2$, then S(f) can, after evaluating f at a suitable finite number of points, always be restricted to a convex set of arbitrarily small area. We cannot however, prove that the functional is localizable, for the functions $f \in F_2$.

A convex set of small area and large diameter is necessarily long and thin; i.e. the ratio of diameter to width must be large. The assumptions for the class F_3 guarantee that if $f \in F_3$ and S(f) can be restricted to a set of small area, then that set must necessarily have small diameter. It turns out that we get localizability for the class F_3, even without the convexity condition as in F_2. The proof of this was obtained by A.M. Gleason [2], in 1961.

Our result for the class F_1 is summarized by the following lemma.

Lemma 1.7 Let k be an arbitrary positive integer, and let $x_1, x_2, \ldots x_k$, be a set of k distinct points in D. Let f in F_1 be such that $S(f) \neq x_i$, $i = 1,2,\ldots k$. Let H be the set of all points z in D such that the closed line segment from z to S(f) contains an x_i.

Then, if y belongs to both the interior of D and the complement of H, there is a $f \in F_1$ such that

(i) $g(x_i) = f(x_i) \quad i = 1,2,\ldots k,$

(ii) $S(g) = y$.

Proof Let L denote the closed line segment joining y and S(f). Since H is closed relative to D, and L does not meet H, there exists an open rectangle U containing L, and contained in the intersection of the complement of H with the interior of D. Let L' be the line through S(f) perpenducular to L, and let V be that open half space bounded by L' which contains y.

Now, every point x in the closure of $U \cap V$ can be written uniquely in the form

$$x = \theta y + (1 - \theta)x'$$

where $0 \leq \theta \leq 1$ and x' lies on the boundary of $U \cap V$.

Define the function g by the equations

$g(x) = f(x)$ if x belongs to both D and the complement of $U \cap V$

$g(y) = f(S(f)) + 1$

$g(x) = \theta\, g(y) + (1-\theta)f(x')$ if x is in $U \cap V$ and $x = \theta y + (1-\theta)x'$.

It is easy to check that $g \in F_1$, $S(g)=y$ and $g(x_i) = f(x_i)$ $i=1,2,\ldots k$.

Thus if $f \in F_1$, and f is evaluated at any finite number of points in D, S(f) may still be anywhere in D except perhaps on a finite number of line segments.

Before considering the class F_2, we state a theorem on plane convex sets which will be needed later. The theorem was originally proved by A. Winternitz. An account of it is to be found in the book by Yaglom and Boltyanskii [10] on page 36.

Theorem 1.8 Let K be a plane convex figure with area A. Then any line L which passes through the centroid of K divides K into two convex subsets, of areas A_1 and A_2, where

$$\frac{4}{9} A \leq A_1 \leq \frac{5}{9} A.$$

We wish to show that if $f \in F_2$, then the possible position of S(f) can be restricted to a convex set of arbitrarily small area by evaluating f at a suitable finite sequence of points in D. It is enough to show the existence of a constant $c < 1$, such that if A is any convex subset of D, and, for some $f \in F_2$ it is known that $S(f) \in A$, then, by evaluating f at suitable points in A, one can restrict S(f) to a certain half-space and thence to a convex set B, where the area of B is less than c times the area of A.

We use the following notation:

A will denote an arbitrary convex subset of D with positive area.

If x_α, x_β are two points of A, then $[x_\alpha, x_\beta]$ will denote the line segment joining x_α and x_β, including its end points.

If x_α, x_β, and x_γ, are three distinct non-collinear points in the interior of A, we can complete the triangle with vertices at x_α, x_β, and x_γ, and extend all sides to meet the boundary of A. Of the six open regions (open relative to A) exterior to the triangle so formed, $\Gamma_{\alpha\beta\gamma}$ will denote the one whose closure has the vertex x_α, and only the vertex x_α in common with the triangle.

We prove first a lemma to describe how, by evaluating a function $f \in F_2$ at three points, the position of S(f) may be restricted. There follows a lemma to show how S(f) can then be restricted to a suitable half-space.

Lemma 1.9 If $f \in F_2$, A is given, $S(f) \in A$, x_α, x_β, x_γ, are any three distinct non-collinear points in A, and if

$$f(x_\alpha) \leq \text{Min } \{f(x_\beta), f(x_\gamma)\}$$

then $S(f) \notin \Gamma_{\alpha\beta\gamma}$.

Proof If $S(f) \in \Gamma_{\alpha\beta\gamma}$, then there is a point z in $[x_\beta, x_\gamma]$ such that x_α lies in the line segment $[z, S(f)]$. But this implies that

$$f(z) < f(x_\alpha) \leq \text{Min}\{f(x_\beta), f(x_\gamma)\}$$

which contradicts the assumptions concerning F_2 .

Lemma 1.10 Suppose A is given, $f \in F_2$, and $S(f) \in A$. Let ξ be the centroid of A, and let $\epsilon > 0$ be given, ϵ less than half the shortest distance from ξ to the boundary of A. Then there is a strategy with respect to which S(f) can be restricted to a half-plane with boundary tangent to the circle, with centre at ξ and radius 2ϵ .

Proof Select points x_1, x_2, x_3, in A to form the vertices of an equilateral triangle with centre at ξ and edge length less than ϵ.

Evaluate $f(x_i)$, $i = 1, 2, 3$. By relabelling the three points if necessary, it may be assumed that $f(x_3) \leq \text{Min}\{f(x_1), f(x_2)\}$. By lemma 1.9, $S(f) \notin \Gamma_{321}$.

Select points x_4, x_5, in A, such that the points x_1, x_2, x_5, x_4 form the vertices of a rectangle with x_3 as the mid-point of $[x_4, x_5]$. By convexity, at least one of the numbers $f(x_4)$, $f(x_5)$ is less than $f(x_3)$. By relabelling x_4 and x_5 if necessary, it may be assumed that $f(x_4) < f(x_3)$ and by relabelling x_1 and x_2 if necessary, $[x_1, x_4]$ is a side of the rectangle.

One now has $S(f) \notin \Gamma_{413}$, by lemma 1.9, since $f(x_4) < f(x_3) \leq f(x_1)$.

There are two possibilities

(i) $f(x_5) \leq f(x_3)$. Then $S(f) \notin \Gamma_{523}$, and the two regions Γ_{413}, Γ_{523} together define a half-space not containing $S(f)$ which satisfies the requirements of the lemma. In this case the proof is finished.

(ii) $f(x_5) > f(x_3)$. Then $S(f) \notin \Gamma_{325}$, and since also $S(f) \notin \Gamma_{321}$, $S(f) \notin \Gamma_{315}$, which is a region with a vertex angle of 120° at x_3.

Proceeding from case (ii), select two points x_6 and x_7 so that the three points x_6, x_3, and x_7, are collinear, and perpendicular to $[x_2, x_3]$, and such that the triples at points x_1, x_6, x_4 and x_2, x_5, x_7 are collinear. Evaluate $f(x_6)$ and $f(x_7)$, and without loss of generality, assume that $f(x_6) < f(x_3)$. Then $S(f) \notin \Gamma_{617}$ by lemma 1.9.

Again there are two possibilities

(i) $f(x_7) \leq f(x_3)$, from which it follows that $S(f) \notin \Gamma_{723}$, and as in case (i) above the theorem is proved.

(ii) $f(x_7) > f(x_3)$, whence $S(f) \notin \Gamma_{317}$, which is a region with a vertex angle of 150° at x_3.

Clearly, the whole processes can be continued for as long as desired. At each stage, two new points are selected, lying on a line through x_3, which is perpendicular to the bisector of the vertex angle at x_3 of the largest region which is known not to contain $S(f)$. Either we finish at some stage when case (i) applies, or, for some two points x_α and x_β, we have, after n stages involving the evaluation of f at $2n+1$ points in A, $S(f) \notin \Gamma_{3\alpha\beta}$, which is a region with vertex angle at x_3 equal to

$180^{\circ} - (\frac{1}{2})^{n} 240^{\circ}$, $n \geq 1$. Moreover, every point at which f is evaluated is within distance 2ϵ of ξ and so belongs to A.

The line segments $[x_\alpha, x_3]$ and $[x_\beta, x_3]$ can be extended beyond x_3 to meet the boundary of A at two points y_α and y_β respectively. An easy calculation shows that the line segment $[y_\alpha, y_\beta]$ passes within 2ϵ of ξ, provided n is chosen so that $2^{-n} \leq 3\epsilon/(4\pi d)$, where d is the diameter of A. This completes the proof of the lemma.

Theorem 1.11 If A is an arbitrary convex subset of D, and if $f \in F_2$, $S(f) \in A$, then by evaluating f at a suitable sequence of points in A, S(f) can be restricted to a convex set B, $B \supset A$, where the (area of B) $\leq \frac{2}{3}$(area of A).

Proof Let A denote the area of A, and d the diameter of A. Let ξ be the centroid of A, and w the shortest distance from ξ to the boundary of A. If $w = 0$, then $A = 0$ and there is nothing to prove, so without loss of generality, $w > 0$, $A > 0$, $d > 0$.

Choose $\epsilon > 0$, so that $\epsilon < \text{Min} \{ \frac{w}{2}, \frac{A}{18d} \}$, and choose a positive integer n so that

$$2^{-n} < 3\epsilon / (4\pi d).$$

Apply the strategy defined by lemma 1.10. Since $\epsilon < \frac{w}{2}$, all points at which f is evaluated belong to A. Hence S(f) can be restricted to a half-space containing ξ which is bounded by a line tangent to the circle with centre at ξ and radius less than 2ϵ, by evaluating f at 2n+1 suitable points in A. If B is the intersection of A with this half-space, then B is convex, and by theorem 1.8, the area of B is bounded above by $\frac{5}{9} A + 2\epsilon d$. Since $\epsilon < \frac{A}{18d}$, the theorem is proved.

Corollary 1.12 If $f \in F_2$, there exists a strategy such that $S(f)$ can be successively localized to a sequence

$$A_0 = D \supset A_1 \supset A_2 \supset \ldots \supset A_k \supset \ldots$$

of convex subsets of D, where for each positive integer k, the area of A_k is less than or equal to $(\frac{2}{3})^k$.

Remark The choice of the integer n in lemma 1.10 is determined by the inequalities

$$2^{-n} < 3\epsilon \,/\, (4\pi d) \leq A \,/\, (24\pi d^2).$$

This means that n cannot necessarily be a fixed integer, since the longer and thinner A is, the larger n has to be.

Kiefer [6] has shown that F_2 contains a subclass F_2' for which, if $f \in F_2'$, then $S(f)$ cannot be restricted to a convex set of diameter less than 1. Hence, if $f \in F_2$, we cannot expect to do better than to restrict $S(f)$ to a convex set of arbitrarily small area.

We shall not discuss all the details of the proof that the functional is localizable for functions in F_3, since this problem is considered in full by A.M. Gleason [2], for the case where D is the n-dimensional unit cube. Gleason proves the following theorem:

Theorem 1.13 If D is the unit cube in Euclidean n-space, and if $f \in F_3$, then there is a strategy such that for any given positive ϵ, $S(f)$ can be restricted to a ball of radius ϵ, by computing the values of f at, at most, L-K Log ϵ points in D. L and K are constants depending only on σ and on the dimension n.

We give a brief description of the strategy used by Gleason.

Let ρ, $0 < \rho < 1$, be arbitrary. Suppose it is known

that for some point $y \in D$, $\{x \in D \mid f(x) \geq f(y)\} \subset \bar{B}(y,\lambda)$, where $\bar{B}(y,\lambda)$ is the closed ball with centre y and radius λ. Choose points $z_0 = y, z_1, z_2, \ldots z_q$, $q < \infty$, so that

$$D \cap \bar{B}(y,\lambda) \subseteq \bigcup_{i=0}^{q} B(z_i, \rho\sigma\lambda).$$

Compute $f(z_i)$, $i = 0,1,2,\ldots q$, and let s be the least index for which $f(z_s) = \sup \{f(z_i)\}$, $i = 0,1,2,\ldots q$. Gleason proves that then

$$\{x \in D \mid f(x) \geq f(x_s)\} \subset \bar{B}(z_s, \rho\lambda).$$

Thus, knowing that S(f) belongs to a ball of radius λ, we can restrict S(f) to one of radius $\rho\lambda$, and this procedure can be iterated. Since initially $S(f) \in D \subset \bar{B}(y_0, \frac{1}{2}\sqrt{2})$ where y_0 is the centre of D, we can after n stages restrict S(f) to a ball of radius $\rho^n \frac{1}{2}\sqrt{2}$. Since $\rho < 1$, $\lim_{n \to \infty} \rho^n \frac{1}{2}\sqrt{2} = 0$. So the maximum point is localizable.

1 - 6 Remarks

Several questions and possible theorems are suggested by the proceding examples. In the first place, it would seem desirable to find sufficient conditions on D and F to give localizability. One possibility for this is given by the three examples given in section 1-1, theorem 1.1, and theorem 1.13. In all three cases, if it is assumed that S(f) belongs to an interval or ball of diameter d say, we can by evaluating f at a fixed number of points in D restrict S(f) to another interval or ball of diameter d' say, where the ratio d'/d is less than 1 and is independent of the value of d. This not only gives localizability by iteration, but also the fact that the diameters of the successive sets to which S(f) is confined decrease at least as fast as a

geometric sequence of positive numbers converging to zero.

Secondly, in the cases where the functional is not localizable, as in section 1-4 and theorem 1.11, it is by examining the way in which localization fails that suggests the extra restrictions to be placed on F in order to give localizability. This suggests that in order to decide whether or not a functional is localizable, it might be possible to examine the sequence of sets to which S(f), $f\in F$, may be restricted, rather than the class F itself. For, if $f\in F$, and S(f) is successively restricted to the sets $K_0 \supset K_1(f,M) \supset K_2(f,M) \ldots \supset K_n(f,M) \ldots$ where M is some strategy, there are just two possibilities:

(i) The diameters $d_n(f,M)$ of the sets $K_n(f,M)$ decrease to a positive limit as $n \to \infty$.

(ii) The diameters $d_n(f,M)$ decrease to zero.

If we have localizability with respect to M, then case (ii) must hold for all functions $f\in F$. Hence to ask for localizability, is the same thing as to ask if the sequence of sets $\{K_n(f)\}$ converge to S(f) in the sense of filters on a topological space. In chapter 2, we define a weak topology on the set K_0, and characterize localizability in terms of the relation between this weak topology and the standard metric topology on K_0.

Finally, the bisection problem provides an example where, for $f\in F$ and $x\in D$, the full information provided by the numerical value of f(x) is not used. It is only necessary to know whether f(x) is positive, negative or zero. Similarly, in the unimodal problem of theorem 1.1, only the sign of $f(x_1) - f(x_2)$ is used to restrict the position of S(f). In general, it can be asked if

anything is to be gained by using the full information available, namely, the numerical value of $f(x)$, as opposed to the partial information provided by the sign of $f(x)$. This question is considered in chapter 2, when the concept of test-function is introduced. First, motivated by the examples of this chapter, general definitions and notation are set up for the problem of localizability.

CHAPTER 2

LOCALIZABILITY

2 - 1 Basic definitions

Definition 2.1 Let F be a given family of real-valued functions defined on a compact convex set D, where D is contained in a Euclidean q-dimensional space E^q, $q < \infty$. For a positive integer m, and any set $A \subset E^q$, define A^m to be the Cartesian product space $A^m = A \times A \times A \times \ldots \times A$ (m times), $A^m \subset E^{mq}$, with the usual Euclidean metric in E^{mq}.

It is assumed that for each function $f \in F$, and some fixed positive integer m, there is a unique vector $S(f) \in D^m$. The mapping $S: F \rightarrow D^m$ which takes f to its image S(f) is called a functional. Define the set H_o as the range of S; thus $H_o = \{x \in D^m \mid x = S(f) \text{ for some } f \in F\}$. It is assumed that if $f \in F$ and $x \in D$, then f(x) can be computed numerically without error.

Definition 2.2 If H is any subset of D, a k-abscissa set in H (or, if there is no ambiguity, an abscissa set) is an ordered k-tuple $\{x_1, x_2, \ldots, x_k\}$ where $x_i \in H$, $i = 1,2,\ldots k$. It is convenient to denote such an abscissa set by X.

Definition 2.3 If X is the abscissa set $\{x_1, x_2, \ldots x_k\}$ where $x_i \in D$, and if $f \in F$, then f(X) denotes the set of k real numbers $\{f(x_1), f(x_2), \ldots, f(x_k)\}$.

Definition 2.4 Suppose $f \in F$. A search procedure for S(f) is defined by the following sequence of steps:

(i) Select a bounded set $K_o \subset E^{mq}$, which contains H_o. (The reason for not necessarily selecting H_o itself, is that it may be more convenient to use a larger set. For example, H_o may be open, but it may be simpler to use closed sets, i.e., K_o might be chosen as the closure of H_o. In general, K_o will be the smallest convenient set containing H_o. Similarly for the sets $K_1(f,X_o)$, $K_2(f,X_o,X_1)$, etc. below.

(ii) Select a k_o-abscissa set X_o in D, and evaluate $f(X_o)$.

(iii) Put $H_1(f,X_o) = \{x \in K_o \mid x = S(g)$ for some $g \in F$ such that $f(X_o) = g(X_o)\}$.

(i)' Select a set $K_1(f,X_o)$ such that $H_1(f,X_o) \subseteq K_1(f,X_o) \subset K_o$.

(ii)' Select a k_1-abscissa set X_1 in D, and evaluate $f(X_1)$.

(iii)' Put $H_2(f,X_o,X_1) = \{x \in K_1(f,X_o) \mid x = S(g)$ for some $g \in F$ such that $f(X_1) = g(X_1)\}$.

In general, having selected k_i-abscissa sets X_i, $i = 0,1,2,\ldots n-1$, and defined

$H_i(f,X_o,X_1, \ldots X_{i-1})$ for $i = 1,2,\ldots n$,

$K_i(f,X_o,X_1, \ldots X_{i-1})$ for $i = 1,2,\ldots (n-1)$

the procedure continues:

(i) Select a set $K_n(f,X_o,X_1, \ldots X_{n-1})$ such that $H_n(f,X_o,\ldots X_{n-1}) \subset K_n(f,X_o,X_1,\ldots X_{n-1}) \subset K_{n-1}(f,X_o,\ldots X_{n-2})$.

(ii) Select a k_n-abscissa set X_n in D and evaluate $f(X_n)$.

(iii) Put $H_{n+1}(f,X_o,X_1,\ldots X_n)$
$= \{x \in K_n(f,X_o,\ldots X_{n-1}) \mid x = S(g)$ for some $g \in F$ such that $f(X_n) = g(X_n)\}$.

This procedure is to continue for all positive integers n.

Definition 2.5 A strategy for localizing the functional, usually denoted by M, is a rule for successively determining the selection of the positive integers k_n, the k_n-abscissa sets X_n, the set K_o, and the sets $K_n(f,X_o,X_1,\ldots X_{n-1})$, $n \geq 1$. The strategy M is to be stated before any computations are carried out. Given M, the set $K_n(f,X_o,\ldots X_{n-1})$ must be uniquely defined in terms of $H_n(f,X_o,\ldots X_{n-1})$ alone, and furthermore, given M and a set $K_n(f,X_o,X_1,\ldots X_{n-1})$, the k_n-abscissa set X_n must be uniquely defined in terms of $K_n(f,X_o,\ldots X_{n-1})$, the abscissa sets X_i, and numbers $f(X_i)$, $i = 0,1,2,\ldots,(n-1)$.

Remarks

If M is a given strategy, and $f \in F$, we can now write $K_n(f,M)$ in place of $K_n(f,X_o,\ldots X_{n-1})$. If there is no ambiguity, we can abbreviate this further to $K_n(f)$ or even to K_n. It is obvious that if $f \in F$, then $S(f) \in K_n(f,M)$ for all positive integers n. We say that n stages have been completed if we know that $S(f) \in K_n(f)$.

Definition 2.6 Let M be a strategy for localizing the functional. The functional is localizable with respect to M if

$$\lim_{n \to \infty} \sup_{f \in F} d[K_n(f,M)] = 0$$

where $d[K_n(f,M)]$ is the diameter of the set $K_n(f,M)$ in the Euclidean space E^{mq}.

2 - 2 Test functions

For many problems, the search procedure as defined in definition 2.4, may be replaced by something more restrictive but

simpler. One possibility is provided by the bisection problem. Suppose it is already determined that, for some $f \in F$, $S(f) \in K_n$, where K_n is a closed interval $[a_n, b_n]$. If, for the abscissa set X_n, the point $\frac{a_n + b_n}{2}$ is selected, one needs to consider (essentially) only two possible sets for K_{n+1}, namely:

$$K_{n+1} = [a_n, \frac{a_n + b_n}{2}]$$

which is to be used if $f\left(\frac{a_n + b_n}{2}\right) < 0$

$$K_{n+1} = [\frac{a_n + b_n}{2}, b_n]$$

which is to be used if $f\left(\frac{a_n + b_n}{2}\right) > 0$.

We therefore ask if we can find a function T of X_n and $f(X_n)$ alone, with finite range, so that, given $S(f) \in K_n$, the position of S(f) can be suitably restricted to a set $K_{n+1}(f)$ which is completely determined by the set K_n and the value of $T(X_n; f(X_n))$.

Definition 2.7 Let k be a positive integer. A test function of order k is a mapping $T : D^k \times R^k \longrightarrow V$, of the form $T(X; f(X))$, where X is a k-abscissa set in D, $f \in F$, and V is a set with a finite number of elements.

Definition 2.8 Let D and F be given. Let T be a k-th order test function with range $\{\lambda_1, \lambda_2, \ldots \lambda_w\}$. A k-th order search procedure with the test function T is defined by the following sequence of steps:

(i) Select a bounded set K_o, so that $H_o \subset K_o \subset D^m$.

(ii) Select a k-abscissa set X_o in D.

(iii) Define w subsets K_{1i}, $1 \leq i \leq w$, of K_o, depending only on X_o and K_o, such that for every $f \in F$, $S(f) \in K_{1i}$ if

$T(X_o;\ f(X_o)) = \lambda_i,\quad i = 1,2,3,\ldots w.$

(iv) For each set K_{1i}, select a k-abscissa set X_{1i} in D.

(v) For each set K_{1i}, define w subsets K_{2j}, $(i-1)w < j \leq iw$, depending only on X_{1i} and K_{1i}, so that for every $f \in F$ satisfying $S(f) \in K_{1i}$, $S(f) \in K_{2j}$ if $T(X_{1i};\ f(X_{1i})) = \lambda_j$.

In general, having defined sets K_{ni}, $1 \leq i \leq w^n$ for some positive integer n, the procedure is continued by

(i) selecting for each set K_{ni}, a k-abscissa set X_{ni},

(ii) defining, for each set K_{ni}, w subsets $K_{n+1,j}$ of K_{ni}, $(i-1)w < j \leq iw$, depending only on K_{ni} and X_{ni}, so that for every $f \in F$ satisfying $S(f) \in K_{ni}$, $S(f) \in K_{n+1,j}$ if $T(x_{ni}; f(X_{ni})) = \lambda_j$.

This procedure is to be continued for all positive integers n.

Definition 2.9 If D, F and a kth-order test function T are given, a strategy, denoted by M, is a rule by which each k-abscissa set X_{ni} is determined when K_{ni} is known, and by which each $K_{n+1,j}$ is determined when K_{ni} and X_{ni} are known, $(k-1)w < j \leq iw$. The strategy M is to be stated before any computations are carried out.

Definition 2.10 If D and F are given; if T is a kth-order test function, and M is a strategy, then the functional is localizable with respect to T and M if

$$\lim_{n \to \infty} \sup_{1 \leq i \leq w^n} d[K_{ni}] = 0$$

where $d[K_{ni}]$ is the diameter of the set K_{ni} in the Euclidean space E^{mq}.

It is evident that if for some T and M, the functional is localizable with respect to T and M, then the functional is

localizable with respect to some strategy M_1 in the sense of definition 2.6. There are two main questions to consider:

(i) If the functional is localizable with respect to some strategy, does there exist a test function T, and another strategy M so that the functional is localizable with respect to T and M?

(ii) If the answer to (i) is yes, how do the two methods compare?

In chapters 3 and 5 it is shown that for many problems, question (i) can be answered in the affirmative. For the present, we prove below a theorem which gives a necessary condition that, for suitable strategies M, the closure of every set $H_n(f,X_o,\ldots X_{n-1})$ as in definition 2.4 is equal to the closure of some K_{ni} as in definition 2.8. To do this, we must decide first when for two functions f and g in F, and some abscissa set X, we have $T(X;f(X)) = T(X;g(X))$. This is done by partitioning the class F into suitable equivalence classes as in the following definition.

Definition 2.11 Suppose D and F are given, and a suitable set K_o is selected. If $f\in F$, $g\in F$, and X is any k-abscissa set in D, the relation $\underset{X}{\sim}$ is defined by

$$f \underset{X}{\sim} g \quad \text{if and only if } H_1(f,X) = H_1(g,X).$$

It is easy to see that $\underset{X}{\sim}$ is an equivalence relation on F. Hence F is partitioned into a number of equivalence classes 'at X'. Moreover, the set $H_1(f,X)$ depends only on the set X and the equivalence class at X to which f belongs.

Lemma 2.12 Suppose D and F are given. Let M be a strategy for localizing the functional so that each set $K_1(f,X_o)$ is chosen as

the closure of $H_1(f,X_o)$, where $f\in F$, and X_o is an abscissa set in D. Then there is a test function T such that every set $K_1(f,X_o)$ is a K_{1i} as in definition 2.8 if and only if the relation $\widetilde{X_o}$ partitions F at X_o into a finite number of equivalence classes. Then T must be defined so that for all $f\in F$, $g\in F$

$$f \widetilde{X_o} g \quad \text{if and only if } T(X_o;f(X_o)) = T(X_o;g(X_o)).$$

Proof The finiteness of the number of equivalence classes is evidently necessary since a test function has finite range. Conversely, if $\widetilde{X_o}$ partitions F into a finite number of equivalence classes, the number of distinct sets $H_1(f,X_o)$ for fixed X_o is finite, and so the number of distinct sets $K_1(f,X_o)$ for fixed X_o is also finite.

Finally, f and g both belong to the same equivalence class at X_o if and only if $H_1(f,X_o) = H_1(g,X_o)$, and, by definition of T, this is if and only if

$$T(X_o;\ f(X_o)) = T(X_o;\ g(X_o)).$$

It is worth noting that to find the set $K_1(f,X_o)$, for some $f\in F$, one need only compute $T(X_o;f(X_o))$, and that this need not entail evaluating all the numerical values of $f(X_o)$. Of course, it does not necessarily follow that with T defined as in lemma 2.12, the functional is localizable with respect to T and M.

2 - 3 Examples

Example 2.13 We take the bisection problem as defined in section 1-1, with domain D = [0,1]. Choose for the strategy M, the one defined by the following rules:

(i) Every set K_n is the closure of the corresponding H_n.

(ii) Given K_n, the abscissa set X_n is the mid-point of K_n.

It follows that the first abscissa set X_o is $x_o = \frac{1}{2}$, and that $K_1(f,x_o) = [x_o,1]$ if and only if $f(x_o) > 0$

$K_1(f,x_o) = [0,x_o]$ if and only if $f(x_o) < 0$

$K_1(f,x_o) = x_o$ if and only if $f(x_o) = 0$.

Hence, if $f \in F$, $g \in F$,

$$f \underset{x_o}{\sim} g \quad \text{if and only if sgn } f(x_o) = \text{sgn } g(x_o).$$

The test function T can be written in the form $T(x,f(x)) = \text{sgn } f(x)$; the three values of which at $x=x_o$ determine uniquely the three sets $K_1(f,x_o)$. Since using the test function T gives the same results as in section 1-1, nothing is to be gained by computing the numerical value of $f(x)$, at any point x in D.

Example 2.14 For the unimodal problem, as in section 1-3, an abscissa set must consist of at least two points in D. If X_o is the pair of points $\{x_1,x_2\}$ with $x_1 < x_2$, and if f and g are two functions in F, then

$$f \underset{X_o}{\sim} g \text{ if and only if } \text{sgn}[f(x_1)-f(x_2)] = \text{sgn}[g(x_1)-g(x_2)]$$

So the test function T can be written

$$T(x;f(x)) = T(x_1,x_2;f(x_1),f(x_2)) = \text{sgn}[f(x_1)-f(x_2)].$$

Example 2.15 Consider the problem in theorem 1.13; to localize the maximum point of a function $f \in F_3$, defined on the unit square D. For certain constants ρ, σ, and λ, one takes as an abscissa set X_o in D a set of q+1 distinct points $z_o,z_1,\ldots,z_q$, so that the union of the open balls with centre z_i and radius $\rho\sigma\lambda$ contains

D. The class F_3 has the property that if $f \in F_3$, and $f(z_i)$ are all computed, then $S(f)$ belongs to a certain ball with centre at z_s, where s is the least index such that

$$f(z_s) = \sup \{f(z_i),\ i = 0,1,2,\ldots q\}.$$

Thus, for given X_o, and any two functions f and g in F_3,

$$f \underset{X_o}{\sim} g \quad \text{if and only if both } f(z_s) \geq f(z_i) \quad 0 \leq i \leq q$$

$$g(z_s) \geq g(z_i) \quad 0 \leq i \leq q$$

where s is minimal in both cases. The test function T must have a range consisting of q+1 elements, one for each point z_i.

Finally, we give an example where a test function can be used, but only by failing to use all the available information.

Example 2.16 Let $D = [0,1]$ and F the class of all convex monotone functions f defined on D for which $f(0) = 1$ and $f(1) = -1$. By convexity, it is meant that if $f \in F$, and $0 \leq x_1 < x_2 < x_3 \leq 1$, then $f(x_2) \leq h$, where the points $(x_1, f(x_1))$, (x_2, h) and $(x_3, f(x_3))$ are collinear. If $f \in F$, let $S(f)$ be the (unique) zero of f.

It is obvious that localization is possible by the bisection strategy.

However, as described by Gross and Johnson [3] or Bellman and Dreyfus [1], one can do much better. For,

$$\begin{aligned} H_o &= \{y \in D \mid y = S(g) \text{ for some } g \in F\} \\ &= \{y \in D \mid 0 < y \leq \frac{f(0)}{f(0) - f(1)}\} \\ &= \{y \in D \mid 0 < y \leq \tfrac{1}{2}\}. \end{aligned}$$

If one takes for K_o the closure of H_o, and selects a single point x in K_o as the first abscissa set, then the closure of

$H_1(f,x)$ is

$$\begin{cases} [\xi, \eta] & \text{if } f(x) > 0 \\ x & \text{if } f(x) = 0 \\ [\text{Max}\{0, \eta\}, \xi] & \text{if } f(x) < 0 \end{cases}$$

where $\xi = x/(1-f(x))$ and $\eta = (f(x) + x)/(f(x) + 1)$.

Hence for given x, the length of the interval $H_1(f,x)$ varies continuously with the numerical value of $f(x)$. So no test function can be used with range elements which describe each set $H_1(f,x)$.

2 - 4 L-topology

We turn now to the question of defining the topology as indicated at the end of chapter 1. Let us suppose that the domain D and the class F are given, along with the set $H_o \subset D^m$ for some known positive integer m. Let M be a strategy for which K_o is the closure of H_o. Whether the functional is localizable with respect to M or not, there is always defined for each $f \in F$, a decreasing sequence of sets $K_o, K_1(f,M), K_2(f,M) \ldots K_n(f,M) \ldots$, such that for all positive integers n, $S(f) \in K_n(f,M)$ and so $S(f) \in \bigcap_{k=1}^{n} K_k(f,M)$. The question of localizability depends mostly on the behaviour of $\bigcap_{k=1}^{n} K_k(f,M)$ as n increases, and this is suggestive of the theory of filters on a topological space.

In what follows, we mean by the metric topology on K_o the relative topology on K_o induced by the Euclidean metric of the smallest Euclidean space containing K_o. The terms 'open' and 'closed' etc., are relative to this topology. It is convenient to restate the definition of localizability in another form.

Lemma 2.17 Let M be a strategy such that K_o is the closure of H_o. Then the functional is localizable with respect to M if and only if for every $\epsilon > 0$, there is an integer n_o such that for all x in K_o,

$$K_n(f,M) \subseteq K_o \cap B(x,\epsilon) \text{ if } n \geq n_o, f \in F, \text{ and } S(f) \in B(x,\tfrac{\epsilon}{2}),$$

where $B(x,\epsilon)$ is the open ball with centre x and radius ϵ.

The proof is omitted, since it follows directly from definition 2.6.

We can now define a topology on the set K_o in terms of the class F and the strategy M.

Definition 2.18 Suppose M is given. A subset V of K_o will be called L-open (with respect to M) if and only if V is open (metric), and for every $x \in V$, there exists an $\epsilon_o > 0$ and a positive integer n_o such that $K_n(f,M) \subset V$ if $n \geq n_o, f \in F$, and $S(f) \in B(x,\epsilon_o) \subset V$.

Lemma 2.19 The class of all subsets of K_o which are L-open are the open sets of a topology for K_o. This topology is called the L-topology with respect to M for K_o.

Proof We have to show that the class of all L-open sets is closed under arbitrary unions and finite intersections. For this purpose, let I be an indexing set, and $\{V_\alpha\}_{\alpha \in I}$ be a collection of L-open subsets of K_o. If $x \in \bigcup_{\alpha \in I} V_\alpha$, then $x \in V_\gamma$ for some $\gamma \in I$. So there is an $\epsilon_o > 0$ and an integer n_o, such that $K_n(f,M) \subset V_\gamma$ if $n \geq n_o$, $f \in F$, and $S(f) \in B(x,\epsilon_o) \subset V$, since V_γ is L-open. But then $K_n(f,M) \subset \bigcup_{\alpha \in I} V_\alpha$ and so $\bigcup_{\alpha \in I} V_\alpha$ is L-open.

If V_1 and V_2 are L-open subsets of K_o, and $x \in V_1 \cap V_2$, then there is a δ such that $B(x,\delta) \subset V_1 \cap V_2$. Since $x \in V_1$, there is an $\epsilon_1 > 0$ and an integer n_1 such that $K_n(f,M) \subset V_1$ if $n \geq n_1$

and $S(f) \in B(x, \epsilon_1) \subset V_1$. Since $x \in V_2$, there is an $\epsilon_2 > 0$ and an integer n_2 such that $K_n(f,M) \subset V_2$ if $n \geq n_2$ and $S(f) \in B(x, \epsilon_2) \subset V_2$. If $\epsilon_o = \text{Min } \{\delta, \epsilon_1, \epsilon_2\}$ and $n_o = \text{Max } \{n_1, n_2\}$, then $K_n(f,M) \subset V_1 \cap V_2$ if $n \geq n_o$, $f \in F$, and $S(f) \in B(x, \epsilon_o) \subset V_1 \cap V_2$. By induction, the intersection of finitely many L-open sets is again L-open.

Finally, the set K_o itself is L-open. This completes the proof.

We can now characterize localizability in terms of the L-topology.

Theorem 2.20 Let D, F and a strategy M be given such that K_o is the closure of H_o. Then the functional is localizable with respect to M if and only if the L-topology with respect to M is stronger than the metric topology for K_o. Alternatively this is if and only if every open set (metric) in K_o is L-open.

Proof Suppose V is open (metric) in K_o, $x \in V$, and the functional is localizable with respect to M. Then there is an $\epsilon_o > 0$ such that $B(x, 2\epsilon_o) \subset V$. But then by lemma 2.17, there is an integer n_o such that $K_n(f,M) \subset K_o \cap B(x, 2\epsilon_o) \subset V$ if $n \geq n_o$, $f \in F$, and $S(f) \in B(x, \epsilon_o) \subset V$. So V is L-open.

Conversely, let $\epsilon > 0$ be arbitrary, and suppose that every open set is L-open. It must be shown that there is an integer n_o, such that for every $x \in K_o$, $K_n(f,M) \subset K_o \cap B(x, \epsilon)$ if $n \geq n_o$, $f \in F$, and $S(f) \in B(x, \frac{\epsilon}{2})$. By hypothesis, $K_o \cap B(x, \frac{\epsilon}{2})$ is L-open for every $x \in K_o$. So if x is given, there is a number $\epsilon_o(x) < \frac{\epsilon}{2}$, and an integer $n_o(x)$ such that $K_n(f,M) \subset K_o \cap B(x, \frac{\epsilon}{2})$. Now the union of all open balls $B(x, \epsilon_o(x))$ covers K_o, and since K_o is compact, a finite subcover $B(y_i, \epsilon_o(y_i))$ where $y_i \in K$, $i = 1, 2, \ldots p$, for K_o can be selected, such that

$$K_o \subset \bigcup_{i=1}^{p} B(y_i, \epsilon_o(y_i)).$$

Let $n_o = \text{Max } \{n_o(y_i), n_o(y_2), \ldots n_o(y_p)\}$.

Let $x \in K_o$, and suppose that $f \in F$, $S(f) \in B(x, \frac{\epsilon}{2})$. Then $S(f) \in B(y_i, \epsilon_o(y_i))$ for some i, and so

$$K_n(f,M) \subset K_o \cap B(y_i, \tfrac{\epsilon}{2}) \quad \text{if } n \geq n_o,$$

which gives

$$K_n(f,M) \subset K_o \cap B(x,\epsilon) \quad \text{if } n \geq n_o, \ f \in F, \text{ and } S(f) \in B(x, \tfrac{\epsilon}{2}).$$

Remark Minor modifications must be made in the above definitions and results if one considers 'non-uniform' localizability, meaning that definition 2.6 is replaced by the requirement that the functional is localizable with respect to M is and only if for every $f \in F$, $\lim_{n \to \infty} d[K_n(f,M)] = 0$. The statement of theorem 2.20 remains valid if the L-topology is defined on the set H_o (rather than K_o). Here, a subset V of H_o is L-open if and only if for every $x \in V$, there is an integer n_o such that if $S(f) = x$, and $n \geq n_o$, then $K_n(f,M) \subset V$. Under this definitions, there may be L-open sets which are not (metric) open, as in the bisection strategy with x as the mid-point of D. The point x is L-open, but not open.

Example 2.21 As an application, we consider the problem defined in section 1-4. There, it was shown that the functional is localizable if for all $f \in F$, $S(f)$ is bounded away from an end of the interval D, but not if $S(f)$ may be arbitrarily close to either end of the interval. Since all L-open sets are (metric) open by definition, it follows that for the strategy defined at the end of 1-4, the L-topology for $K_o = [0,1]$, is that with the following sets as a basis:

(i) Intervals of the form $\{x|a < x < b\}$, where $0 < a < b < 1$.

(ii) Unions of disjoint intervals of the form $\{x|0 \leq x < a\} \cup \{x|b < x \leq 1\}$.

For no strategy M, can all sets of the form $\{x|0 \leq x < a\}$ be L-open.

Example 2.22 The L-topology for the class F_1 defined in section 1-5 is the indiscrete topology for the unit square. That is, only the unit square and the empty set are L-open.

2 - 5 A simultaneous search problem

As a further example to illustrate how the L-topology describes how localization may fail, we consider a problem where two points in D are to be simultaneously approximated by localization. Aside from the topology, this problem is of some interest in its own right.

Let D be the closed interval $[0,1]$. Let F be the class of real-valued continuous functions f defined on D such that for two distinct points $\alpha_1(f)$ and $\alpha_2(f)$ in D, $0 < \alpha_1(f) < \alpha_2(f) < 1$, f is strictly increasing on $[0,\alpha_1(f)]$, f is strictly decreasing on $[\alpha_1(f),\alpha_2(f)]$, f is strictly increasing on $[\alpha_2(f),1]$. Let S(f) be the vector $(a_1(f), \alpha_2(f))$ in the plane E^2. We take as K_o, the set $K_o = \{(x,y)|0 \leq x \leq y \leq 1\} \subset E^2$.

We show first that the functional is not localizable. Intuitively, this is because F contains functions f which can be regarded as increasing on D except for a small 'wiggle', which may be contained in an interval so small that one cannot be sure of finding it for any given strategy.

Lemma 2.23 Let k be an arbitrary positive integer, and let $x_1, x_2, \ldots x_{k+1}$ be any set of (k+1) distinct points in D such that

$$0 \leq x_1 < x_2 < x_3 \ldots < x_{k+1} \leq 1.$$

Then for each t, t = 1,2,... k, and every pair of real numbers $\xi = \xi(t)$, $\eta = \eta(t)$ such that $x_t < x_t + \xi < x_t + \eta < x_{t+1}$, there is a function $f_{t,\xi,\eta} \in F$, such that

(i) $f_{t,\xi,\eta}(x_i) = x_i \quad i = 1,2,\ldots k+1$

(ii) $S(f_{t,\xi,\eta}) = (\xi(t), \eta(t))$.

Proof Define $f_{t,\xi,\eta}(x)$ such that

$f_{t,\xi,\eta}(x) = x$ if $0 \leq x \leq t$ or $x_{t+1} \leq x \leq 1$

$f_{t,\xi,\eta}(\xi) = \xi + 1$

$f_{t,\xi,\eta}(\eta) = \eta - 1$

$f_{t,\xi,\eta}$ is linear on the intervals $[x_t, \xi]$, $[\xi, \eta]$ and $[\eta, x_{t+1}]$.

Since $d[H_1(f_{t,\xi,\eta}, X_o)] = \sqrt{2} = d[K_o]$, where X_o is the (k+1) abscissa set $\{x_1, x_2 \ldots x_{k+1}\}$, the functional is not localizable.

One suspects from the above lemma that localization is possible if the two points $\alpha_1(f)$, $\alpha_2(f)$ are uniformly bounded away from each other. This is indeed the case. Suppose for the moment, that for given $f \in F$, a point R(f), such that $\alpha_1(f) < R(f) < \alpha_2(f)$, can be found. Then the function f, restricted to the interval [0,R(f)] is unimodal, with a unique maximum point at $\alpha_1(f)$, and so $\alpha_1(f)$ can be localized as described in section 1-3. Similarly, one can localize $\alpha_2(f)$, and hence S(f).

Two restrictions on the class F are considered which guarantee that for each f, a point R(f), $\alpha_1(f) < R(f) < \alpha_2(f)$ can be found. The first is to take, for some positive number δ the class F_δ which consists of all functions $f \in F$ for which $\alpha_2(f) - \alpha_1(f) \geq \delta$. For each $\delta > 0$, the functional defined by F_δ is localizable. This is easily seen: choose as abscissa set $X_o = \{x_1, x_2, \ldots x_{k+1}\}$ so that

(2.1) $\quad 0 = x_1 < x_2 < x_3 \ldots < x_{k+1} = 1$, where

$$\sup_{i=1}^{k} (x_{i+1} - x_i) \leq \frac{\delta}{4} .$$

Then, since f is strictly decreasing over an interval of length δ, if $f \in F_\delta$, there is an index t, $2 \leq t \leq k$, such that $f(x_{t-1}) > f(x_t) > f(x_{t+1})$. But then, $\alpha_1(f) < x_t < \alpha_2(f)$, and so, by the above remarks, the functional is localizable for functions $f \in F_\delta$.

As a corollary, if we put $K_o(\delta)$ equal to the closure of H_o for the functions $f \in F_\delta$, then $K_o(\delta)$ is the intersection of $\{(x,y) \mid 0 \leq x < y \leq 1\}$ with $\{(x,y) \mid y \geq x + \delta\}$, and there is a strategy M for which the L-topology for $K_o(\delta)$ is exactly the relative metric topology for $K_o(\delta)$.

Suppose such a δ is not known. Then one can attempt to localize S(f) for $f \in F$, by assuming $f \in F_{\delta_o}$, for some δ_o. If S(f) cannot be restricted, by selecting the abscissa set (2.1) appropriate for $\delta = \delta_o$, the process can be repeated for $\delta_1 = \frac{1}{2}\delta_o$, and if necessary, for $\delta_2, \delta_3 \ldots$ where $\delta_n = (\frac{1}{2})^n \delta_o$. This strategy, it being understood that as soon as a suitable δ_n is found, one localizes $\alpha_1(f)$ and $\alpha_2(f)$, has as L-topology that defined by the basis consisting of

(i) K_o intersected with open balls whose closure does not contain a point of the line y=x,

(ii) strips of the form $K_o \cap \{(x,y) | x \leq y < x+\delta\}$.

Every L-neighbourhood of a point on the line y=x contains every other point on that line in K_o.

An alternative restriction on the class F to give localizability lies in the possibility that one can approximate an inflexion point R(f), $\alpha_1(f) < R(f) < \alpha_2(f)$. The restrictions placed on F so that this can be done are analogous to those in the unimodal problem, except that convexity conditions replace monotonicity ones, and the second-order difference quotient

$$\left(\frac{f(x_3) - f(x_2)}{x_3 - x_2} - \frac{f(x_2) - f(x_1)}{x_2 - x_1} \right) \Big/ (x_3 - x_1)$$

is used, rather than the first order difference quotient

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} .$$

The exact conditions on F to give localizability of the functional defined by S(f) are given in theorem 2.25 below; first we define the term 'strictly convex'.

Definition 2.24 Suppose f is a real-valued function defined on the closed interval [a,b]. Then f is strictly convex up on [a,b] if, for all points x_1, x_2, x_3 such that $a \leq x_1 < x_2 < x_3 \leq b$, $(x_3 - x_1)f(x_2) < (x_2 - x_1)f(x_3) + (x_3 - x_2)f(x_1)$.

f is strictly convex down if, under the same conditions, $(x_3 - x_1)f(x_2) > (x_2 - x_1)f(x_3) + (x_3 - x_2)f(x_1)$.

Theorem 2.25 Let D be the closed interval [0,1], and F the class of all real continuous functions f defined on D for which there are three numbers $\alpha_1(f)$, R(f), $\alpha_2(f)$ such that

$0 < \alpha_1(f) < R(f) < \alpha_2(f) < 1$, and

f is strictly increasing on $[0, \alpha_1(f)]$

f is strictly decreasing on $[\alpha_1(f), \alpha_2(f)]$

f is strictly increasing on $[\alpha_2(f), 1]$

f is strictly convex up on $[0, R(f)]$

f is strictly convex down on $[R(f),1]$.

Let $S(f)$ be the point $(\alpha_1(f), \alpha_2(f)) \in E^2$. Then the functional defined by $S(f)$ is localizable.

Proof First, the functional defined by $R(f)$ is localizable with respect to the test function T given by

$$T = T(x_1,x_2,x_3;\ f(x_1),f(x_2),f(x_3))$$
$$= (x_3 - x_1)f(x_2) - (x_2 - x_1)f(x_3) - (x_3 - x_2)f(x_1)$$

where $0 \leq x_1 \leq x_2 \leq x_3 \leq 1$.

For, it follows from definition 2.23, that if $0 < x_1 < x_2 < x_3 < 1$, then

$$R(f) \geq x_1 \quad \text{if} \quad T > 0$$
$$R(f) \leq x_3 \quad \text{if} \quad T < 0$$
$$x_1 \leq R(f) \leq x_3 \quad \text{if} \quad T = 0 .$$

Thus, if $f \in F$, $R(f)$ can, by evaluating f at 3 points of D, be restricted to one of the two intervals $[0,x_3]$ or $[x_1,1]$, and by suitable choice of x_1,x_2,x_3 both these intervals can be chosen to have length not greater than $\frac{2}{3}$ that of D.

Hence by iteration, one can by evaluating f at a suitable sequence of 3n points in D, restrict $R(f)$ to an interval of length $(\frac{2}{3})^n$.

A more efficient strategy for localizing $R(f)$ is given in chapter 5.

Secondly, the functional defined by $S(f)$ is localizable. The strategy commences by restricting $R(f)$ to a subinterval of D of length 2c and mid-point ξ, where $0 < c \leq \text{Min}\{\xi, 1-\xi\}$, for

$f \in F$, and c arbitrary. A choice of c is made later, but assume that $c \leq \frac{1}{4}$. Thus $0 \leq \xi - c \leq R(f) \leq \xi + c < 1$.

There are two possibilities:

(a) $\xi \geq \frac{1}{2}$. Then $c \leq \frac{1}{2}\xi$, so two points y_1 and y_2 can be selected such that

$$0 \leq \xi - 2c < y_1 < y_2 < \xi - c \leq R(f).$$

Evaluate $f(y_1)$ and $f(y_2)$. Again, there are two possibilities:

(i) $f(y_1) \geq f(y_2)$. Then $\alpha_1(f) \leq y_2 < R(f)$. This means that a point y_2 such that $\alpha_1(f) \leq y_2 < \alpha_2(f)$ is known. Since f restricted to $[0,y_2]$ is a unimodal function with a unique maximum point at $\alpha_1(f)$, $\alpha_1(f)$ and similarly $\alpha_2(f)$ can be localized (separately) by the golden-mean strategy, and so in this case S(f) can be localized.

(ii) $f(y_1) < f(y_2)$. Then $\xi - 2c < y_1 < \alpha_1(f) < R(f) < \alpha_2(f) < 1$. At worst, the whole procedure can be started again on the interval $[y_1,1]$.

(b) $\xi \leq \frac{1}{2}$. Then $c \leq \frac{1}{2}(1 - \xi)$, and two points z_1 and z_2 can be selected so that

$$R(f) \leq \xi + c < z_1 < z_2 < \xi + 2c \leq 1.$$

By evaluating $f(z_1)$ and $f(z_2)$, there are the two possibilities:

(i) $f(z_1) \geq f(z_2)$, from which one gets $\alpha_1(f) < R(f) < z_1 \leq \alpha_2(f)$, and localization proceeds as in case (i) above.

(ii) $f(z_1) < f(x_2)$, which gives

$$0 < \alpha_1(f) < R(f) < \alpha_2(f) < z_2 < \xi + 2c.$$

With the choice $c = \frac{1}{20}$ of the length of D, the problem can be recommenced on an interval of length at most $\frac{3}{5}$ that of D,

should case (ii) arise. This can be iterated as often as necessary until eventually either case (i) arises, or it is known that $\alpha_1(f)$, $R(f)$, and $\alpha_2(f)$ all lie in an interval of length at most $(\frac{3}{5})^m$ for some sufficiently large integer m. This completes the proof.

There are many problems similar to the one just discussed which can be handled in the same way. One of importance is where $D = [0,1]$, and F is the class of all continuous real-valued functions f defined on D for which $f(0) > 0$, $f(1) > 0$, f vanishes at exactly two points $\alpha_1(f)$ and $\alpha_2(f)$, $0 < \alpha_1(f) < \alpha_2(f) < 1$, and $f(x)$ is negative if $\alpha_1(f) < x < \alpha_2(f)$. As in the previous example, restrictions on F are required before localization is possible.

CHAPTER 3

RATE OF LOCALIZATION

3 - 1 Speed

This chapter is concerned with the rate of decrease of the sets K_o, $K_1(f)$, $K_2(f)$ ...; i.e., with the speed at which the diameters $d[K_n(f)]$ converge. In general terms, suppose that the domain D and the class of functions F are given. Let M be a strategy. Then for each $f \in F$, S(f) is successively restricted to the sets $K_o \supset K_1(f,M) \supset K_2(f,M) \supset \ldots \supset K_n(f,M) \ldots$ where K_o is determined by M and each set $K_n(f,M)$ is uniquely determined by f and M.

Define $d_o = d[K_o]$ to be the diameter of K_o, which is finite since K_o is bounded, and put $d[K_n(f,M)]$ equal to the diameter of the set $K_n(f,M)$, for each positive integer n.

In chapter one, the sequence $d_n'(M) = \sup_{f \in F} d[K_n(f,M)]$ was defined. The significance of this number, for given n, is that by using the strategy M, it is guaranteed that for any f in F, S(f) can be restricted to a set with diameter at most $d_n'(M)$ after the selection of a sequence of n suitable abscissa sets. However, it seem more appropriate to consider the number of points in D at which a function $f \in F$ is evaluated, rather than the number of abscissa sets. This is because different strategies may involve different numbers of points in D for each abscissa set.

Hence we adopt the following definition:

Definition 3.1 Let N be a positive integer, and M a given strategy. For each $f \in F$, let n be the largest positive integer such that if the strategy M is used to search for S(f), then f is evaluated at no more than N distinct points in D.

Define $$d_N(M) = \sup_{f \in F} d[K_n(f,M)] \qquad N \geq 1$$

and $$d_N = \inf_M d_N(M) \qquad N \geq 1 .$$

It is evident from definition 3.1 that $d_N(M) \geq d_{N+1}(M)$, and $d_N \geq d_{N+1}$ for all integers $N \geq 0$. Moreover, $n \leq N$, since each time a new abscissa set X_n is selected, the function $f \in F$ is evaluated at at least one new point in the domain D. From this fact, and from definition 2.6, it follows that if the functional is localizable with respect to M, then $\lim_{N \to \infty} d_N(M) = 0$. Conversely, if M is a strategy such that the number of elements in each abscissa set is uniformly bounded, and if $\lim_{N \to \infty} d_N(M) = 0$, then $\lim_{n \to \infty} \sup_{f \in F} d[K_n(f,M)] = 0$, i.e., the functional is localizable with respect to M.

We are concerned with the behaviour of the ratios $d_N(M)/d_o$ and d_N/d_o as $N \to \infty$. For many problems these sequences are bounded both from above and below by geometric sequences of positive numbers which converge to zero. In terms of numerical analysis, we might say that in these cases, localization is a first-order process. That this is normally to be expected is demonstrated by some of the examples given in chapters one and two.

For instance, in the bisection problem of section 1-1, the domain D = [0,1] may equally well be replaced by any other compact interval [a,b]. One can still restrict the possible

position of S(f) to a subinterval of length $\frac{1}{2}$ that of [a,b]. Since the factor $\frac{1}{2}$ is independent of a and b, the procedure can be iterated in such a way that $d_N(M)/d_o \leq (\frac{1}{2})^N$. Similarly for the unimodal problem, if D = [a,b] is any compact interval on the line, S(f) can be restricted to a subinterval of D of length at most $\frac{3}{4}$ that of D, by evaluating f at at most 2 points in D. Again, the factor $\frac{3}{4}$ is independent of a and b, so that by iteration, there is a strategy M for which $d_{2N}(M)/d_o \leq (\frac{3}{4})^N$ for all positive integers N.

These considerations lead immediately to the following lemma.

Lemma 3.2 Let the domain D and class of functions F be given, with $D \subset E^q$ for some q. Suppose $H_o \subset D^m$ for some positive integer m. A sufficient condition that the functional be localizable is that there exists a positive integer k, a real number c, $0 < c < 1$, and a class B of bounded subsets of E^{mq} such that

(i) $K_o \in B$

(ii) For every set $K \in B$, there exist k points $x_1, x_2, \ldots x_k$, such that if $f \in F$ and $S(f) \in K$, then

$$K \cap \{y \in K_o \mid y = S(g) \text{ for } g \in F \text{ such that } g(x_i) = f(x_i)$$
$$i = 1, 2, \ldots k\} \subset K',$$

where $K' \in B$ and $d[K'] \leq cd[K]$. Moreover, there is a strategy M such that $d_N(M) \leq (\sqrt[k]{c})^N d_o/c$, for all positive integers N.

Proof It is evident that if M denotes the strategy which defines the given k points $x_1, x_2, \ldots x_k$, in terms of K, and restricts S(f) to the corresponding set K', then $d_{pk}(M) \leq c^p d_o$, for all

positive integers p. If N is any positive integer, then there is an integer p such that $pk \leq N \leq pk + k$; so

$$\begin{aligned} d_N(M) &\leq d_{pk}(M) \\ &\leq c^p d_o \\ &= (\sqrt[k]{c})^{pk+k} \; d_o/c \\ &\leq (\sqrt[k]{c})^N \; d_o/c \; . \end{aligned}$$

In particular, the functional is localizable with respect to M.

The class B of lemma 3.2 is more or less arbitrary, though there is usually a simple and convenient choice for any given problem. It is often possible to take for B the class of all homothetic images of K_o; thus, if K_o is a compact interval, so are all the members of B. Occasionally, it may be desirable, especially if the dimension of D or K_o is greater than 1, to choose B as the class of all compact convex subsets of K_o.

We are interested in proving a converse result to lemma 3.2; namely that if a functional is localizable, then it is localizable at at least a geometric rate. This requires a restriction which is perhaps best stated as a separate hypothesis.

Definition 3.3 Suppose the domain D and class F are given, where D is a compact cube in E^q, and $K_o = D$. Consider any compact cube H contained in D. The coordinates of H can be identified with those of D by a mapping θ_H of H onto D consisting of a proper Euclidean motion followed by a change of scale. The class F is called homogeneous with respect to the functional S, if, given any $f \in F$, and a compact subset H of D, containing S(f),

the function f restricted to H becomes a new function $\hat{f}$ defined on D after applying the map θ_H, where $\hat{f}$ satisfies

$$\text{(i)} \quad S(\hat{f}) = \theta\,(S(f))$$
$$\text{(ii)} \quad \hat{f} \in F.$$

Theorem 3.4 If F and D are given, $D = K_o$ being a compact cube in E^q, and if F is homogeneous with respect to the functional S, then all the hypotheses of lemma 3.2 hold. In particular, there is a constant $c < 1$, and a strategy M, such that $d_N(M) \leq (k\sqrt{c})^N d_o/c$, for some k.

Proof Let B be the class of all compact subcubes of D. Let c, $0 < c < 1$ be arbitrary. Since the functional is localizable, there is an integer k, and an abscissa set X_o, such that for all $f \in F$

$$d[H_1(f,X_o)] \leq cd[K_o]\, 2/\sqrt{q}\ .$$

But then, $H_1(f,X_o)$ can be contained in a cube $K_1(f) \in B$ where $d[K_1(f)] \leq cd[K_o]$. The hypotheses guarantee that this can be iterated indefinitely by suitable changes of scale.

Theorem 3.4 can be extended to the general case where for all $f \in F$, $S(f) \in K_o = D^m$ for some integer $m > 1$, D being a compact cube in E^q.

First, definition 3.3 is adopted, replacing the point $S(f) \in H$ by its corresponding m-tuple $[\alpha_1(f), \alpha_2(f), \ldots \alpha_m(f)]$ where $\alpha_i(f) \in D$. Theorem 3.4 then carries over for some constant c_1 for at least a finite number of stages. (K_o, and $H_1(f,X_o)$ must be replaced by D and a subcube of D respectively). The procedure continues indefinitely should $\alpha_1(f)$, ... $\alpha_m(f)$ happen to coincide, otherwise it ceases, for some cube $U \neq D$, when at least two points are near opposite ends of a

diameter of H. If now, H can be split into a finite number of subcubes, each containing a known subset of the points $\alpha_1(f)$, ... $\alpha_m(f)$, and if each of these define localization problems with similar homogeneity conditions as in definition 3.4, localization may proceed, with new constants c for each subset. By continuing in this way, one eventually arrives at the situation where each $\alpha_i(f)$ is known to lie in a cube $H_i(f)$, the $H_i(f)$ being disjoint. If each $\alpha_i(f)$ can be localized as a point in its corresponding cube $H_i(f)$, in the class of functions of the type f restricted to $H_i(f)$, and if again the homogeneity requirement is met, each $\alpha_i(f)$ can be localized at a geometric rate.

If c is now taken as the supremum of all the constants c_1, ... used in the above problems, then $c < 1$, and $d_N(M)$, for some M, is bounded above by a geometric sequence converging to zero.

An example where the above argument works is provided by that in section 2-5.

Quite often, the requirement that D be a cube can be dropped. The only reason for choosing D as a cube is the fact that any sufficiently small subset of a cube can be contained in a proper subcube of a cube. This property is not enjoyed by, for example, discs in a plane. However, for some problems, as in theorem 1.13, this requirement is irrelevant.

The existence of a geometric sequence of positive terms which provides a lower bound for the sequence d_N is less obvious even on intuitive grounds than the corresponding result for upper bounds. This is so because, for given $f \in F$, the information known

about f necessarily increases as the strategy chosen proceeds, and conceivably, this could result in a quicker decrease of d_N for large N compared to the decrease for small N.

We have seen earlier in this chapter that if $d_k(M) \leq cd_o$ for some k, some strategy M, and some $c < 1$, then quite often $d_{pk}(M) \leq c^p d_o$ for all positive integers p. The corresponding statement with the inequality signs reversed is not true. For example, in the unimodal problem of theorem 1.1, where $D = [0,1]$, $d_2(M) > \frac{1}{2}$ for all strategies M. But $d_4(M)$, by a suitable choice of M, can be made arbitrarily close of $\frac{1}{5}$. The main reason for this is that the value of $f(x)$, $f \in F$, $x \in D$, can be used for x belonging to two or more abscissa sets.

The following theorem holds in all the examples considered where D contains all points S(f), and the assumptions of definition 3.3 hold. Its proof is derived in a manner similar to the proof of the inequality $L_n \leq L_{n-1} + L_{n-2}$ in theorem 1.1.

Theorem 3.5 Suppose the domain D and class of functions F are given, where $D = K_o$ is a compact cube contained in the Euclidean space E^q. Assume that there exists an integer k and a constant c, $0 < c < 1$ such that for all closed subcubes H of D, the following conditions hold:

(i) There exist k distinct points, $x_1, x_2, \ldots x_k$, belonging to the interior of H such that if $f \in F$ and $S(f) \in H$, then $\{y \in H \mid y = S(g)$ for some g in F such that $f(x_i) = g(x_i)$, $i=1,2,\ldots k\}$ is contained in a subcube at H which does not contain at least one of the points x_i in its interior.

(ii) If $f \in F$ and $S(f) \in H$, then H is identical with $\{y \in H | y = S(g)$ for some $g \in F$ such that $g(x) = f(x)$ for all x in D but not in H$\}$.

(iii) If $X = \{x_1, x_2, \ldots x_k\}$ is any set of k distinct points in H, then there is a function $f \in F$ such that $S(f) \in H$ and the set $\{y \in H |\ y = S(g)$ for some $g \in F$ such that $f(X) = g(X)\}$ contains a cube H', where $d[H'] \geq cd[H]$.

Then, there is a positive constant λ such that $d_N \geq \lambda c^N$ for all positive integers N.

Remark Condition (ii) guarantees that once it is known that $S(f) \in H$, no further restriction on S(f) may be obtained by evaluating f at points exterior to H. Condition (iii) means that the class F contains 'too many' functions for localization to be 'too fast'.

Proof of Theorem 3.5 As in the proof of theorem 1.1, a positive integer n is selected, $n > k$, and the coordinates of D are changed by a proper Euclidean motion and scalar multiplication so that D is a cube of diameter L_n. L_n is defined to be the largest real number, such that, for all $f \in F$, S(f) can be restricted to a cube of diameter 1 by the evaluation of f at at most n points.

It is sufficient to show that $L_{n-1} \geq cL_n$ for all positive integers $n > k$.

Select k distinct points $\{x_1, x_2, \ldots x_k\}$ in the domain D with diameter L_n, and evaluate $f(x_i)$, $i = 1, 2, \ldots k$, for some $f \in F$. Consider the set $D' = \{y \in D |\ y = S(g)$ for some $g \in F$ such that $g(x_i) = f(x_i)$, $i = 1, 2, \ldots k\}$. It follows from conditions (i) and (ii) that, given $S(f) \in D'$, one can restrict S(f) to a cube of

diameter 1, for at least one suitable choice of the abscissa set $\{x_1, x_2, \ldots x_k\}$, by evaluating f at (n-1) points in D'. By condition (iii) D' contains, for some $f \in F$, a cube of diameter at least cL_n.

Hence, $L_{n-1} \geq cL_n$, and this is true for all $n > k$. Since L_k must be finite, the result follows.

Remark Like theorem 3.4, this result can be extended under suitable conditions to the case where $S(f) \in K_o = D^m$ for some $m > 1$.

3 - 2 Speed with Lipschitz Conditions

An important question, which is here left unanswered, arises out of theorem 3.5. If, for given F and D, the sequence d_N can be bounded from below by a geometric sequence of positive terms converging to zero, what restrictions on F are necessary so that d_N cannot be bounded from below in this way ? In other words, what conditions on F give a sequence d_N which converges to zero faster than any geometric sequence ?

Consider for example the problem of localizing the (unique) zero of a real-valued function f defined on an interval. If all that is known about f is the fact that f is continuous and strictly monotonic, vanishing at a single point in the interval, then as in section 1-1, $d_N \geq (\frac{1}{2})^N$. On the other hand, if f is also known to possess a derivative which is bounded away from zero, and a bounded second-derivative, then the classic regula-falsi sequence

$$x_{n+1} = x_n - \Big((x_n - x_{n-1}) f(x_n)\Big) \Big/ \Big(f(x_n) - f(x_{n-1})\Big)$$

with x_0 and x_1 given, and which can be formulated in terms of definition 2.4, gives a sequence for which the ratio $d_{N+1}/\ d_N$ approaches zero as $N \to \infty$.

The one result we have in this direction is that it is not sufficient for faster than geometric convergence to know that f is continuously differentiable with known upper and lower bounds on the derivatives. More precisely:

Theorem 3.6 Let D be the unit interval $[0,1]$, and let F be the class of real-valued continuously differentiable functions f defined on D, for which $f(0) < 0 < f(1)$, $0 < P \leq f'(x) \leq Q < \infty$, where P and Q are two fixed constants, and $f'(x)$ is the derivative of f at any point $x \in D$. Let $\alpha = S(f)$ be the zero of f. Then there exists a strategy M for which

$$d_N(M) = \left[\frac{1}{2} \left(1 - \frac{P}{Q}\right) \right]^N,$$

and moreover this strategy is best possible in the sense that

$$d_N \geq \left[\frac{1}{2} \left(1 - \frac{P}{Q}\right) \right]^N$$

for all positive integers N.

Proof Let $[a,b]$ be any closed sub-interval of D, and suppose for some $f \in F$, it is known that $\alpha = S(f) \in [a,b]$. Select a point $x \in [a,b]$, and evaluate $f(x)$. There are four possibilities for α, given by

(i) $x - \frac{f(x)}{Q} \leq \alpha \leq b$ if $-Q(b-x) \leq f(x) \leq -P(b-x)$

(ii) $x - \frac{f(x)}{Q} \leq \alpha \leq x - \frac{f(x)}{P}$ if $-P(b-x) \leq f(x) \leq 0$

(iii) $x - \frac{f(x)}{P} \leq \alpha \leq x - \frac{f(x)}{Q}$ if $0 \leq f(x) \leq P(x-a)$

(iv) $a \leq \alpha \leq x - \frac{f(x)}{Q}$ if $P(x-a) \leq f(x) \leq Q(x-a)$.

All of these ineuqalities are best possible in the sense that for given f(x), $\{y \mid a \leq y \leq b,\ y = S(g)$ where $g \in F$ and $g(x) = f(x)\}$ contains the interior of the appropriate interval for α.

The lengths of the two intervals in (i) and (ii) are bounded above by $(b-x)(1 - \frac{P}{Q})$, with equality if $f(x) = -P(b-x)$. Similarly, the lengths of the two intervals in (iii) and (iv) are bounded above by $(x-a)(1 - \frac{P}{Q})$ with equality if $f(x)=P(x-a)$.

It follows, that in order to minimize the maximum of the lengths of the four intervals, one must choose $x = \frac{a+b}{2}$, and then, by evaluating f(x), α can be restricted to an interval of length $\frac{1}{2}(b-a)(1 - \frac{P}{Q})$. Hence, by repeated bisection, S(f) can be restricted to an interval of length $[\frac{1}{2}(1 - \frac{P}{Q})]^N$ by evaluating f at a sequence of N points in D.

To complete the proof, it must be shown that $d_N \geq [\frac{1}{2}(1 - \frac{P}{Q})]^N$. This is done by induction on N. If N = 1, a = 0, and b = 1, then the maximum length of the four intervals is greater than or equal to $\frac{1}{2}(1 - \frac{P}{Q})$, with equality only if $x = \frac{1}{2}$ and $f(x) = \pm \frac{1}{2} P$. So $d_1 \geq \frac{1}{2}(1 - \frac{P}{Q})$.

Let F_1 be the subclass of F for which $f \in F_1$ if and only if for some $\epsilon > 0$, $0 < S(f) < \epsilon$, and $f(x) = Px$ if $\epsilon \leq x \leq 1$. Suppose for some $f \in F_1$, and any positive integer N, it is known on only that $0 < S(f) < [\frac{1}{2}(1 - \frac{P}{Q})]^N$, f having been evaluated at N points in D. Then, if ϵ is small enough, $f\left(\frac{1}{2}[\frac{1}{2}(1 - \frac{P}{Q})]^N\right) = \frac{1}{2} P[\frac{1}{2}(1 - \frac{P}{Q})]^N$ and S(f) cannot be restricted to a sub-interval of the open interval $(0, [\frac{1}{2}(1 - \frac{P}{Q})]^{N+1})$. This completes the

induction, and proves that $d_N \geq [\frac{1}{2}(1 - \frac{P}{Q})]^N$ for all positive integers N.

Remark Unlike the bisection problem of section 1-1, the strategy M of theorem 3.6 uses the full information provided by the numerical value of f(x) whenever this is computed. Further, if P=0, or Q is infinite, then $d_N = (\frac{1}{2})^N$ exactly as in section 1-1.

3-3 Exponents

Given a sequence $\{x_n\}$ which converges to the zero α of a function, one defines in numerical analysis, an index as a measure of the speed at which x_n approaches α. It is convenient to introduce here a similar index for comparing different strategies. We call it an exponent.

As indicated in chapter 1, the exponent could reflect many things associated with the problem. We shall, however, allow dependence only on the strategy, the diameter of sets which are known to contain the point S(f), and the number of points in the domain D at which a function f is evaluated.

Of the many possible indices which may be meaningfully defined, we select the following:

Definition 3.7 Let N be a positive integer. Suppose the domain D, and the class of functions F are given. For any strategy M used to search for the functional, put $\rho_N(M) = [d_o/d_N(M)]^{1/N}$, where $d_N(M)$ is as defined in definition 3.1, and d_o is the diameter of K_o.

Define $$\rho_N = \sup_M \rho_N(M) = [d_o / d_N]^{1/N}$$

$$\bar{\rho} = \varliminf_{N \to \infty} \rho_N$$

$$\rho(M) = \varliminf_{N \to \infty} \rho_N(M)$$

$$\rho = \sup_M \rho(M).$$

$\varliminf_{N \to \infty} \rho_N$ denotes the lower limit of ρ_N as $N \to \infty$.

Definition 3.8 The number ρ is called the exponent of the localization problem. A strategy M for which $\rho(M) = \rho$ is called an optimal strategy.

Definition 3.9 If M_1 and M_2 are two strategies for localizing a functional, M_1 is at least as good as M_2 if $\rho(M_1) \geq \rho(M_2)$.

Remark The terminology of definitions 3.8 and 3.9 can equally be applied to $\rho_N(M)$ and ρ_N, for any positive integer N. This is appropriate for the case where it is decided a priori that no further computation is carried out after the function $f \in F$ has been evaluated at N points in the domain D. For given N, a strategy M for which $\rho_N(M) = \rho_N$ is also called optimal.

Lemma 3.10 For any strategy M, there exist constants c_1 and c_2 such that

$$1 < c_2 \leq \rho_N(M) \leq c_1 < \infty$$

for all sufficiently large N, if and only if

$$d_o c_1^{-N} \leq d_N(M) \leq d_o c_2^{-N}$$

for all sufficiently large N.

Thus the quantity $\rho(M)$ is useful only for first order processes.

Lemma 3.11 $\rho = \bar{\rho}$

Proof If M is any strategy for localizing the functional then

$$\rho_N(M) \leq \sup_M \rho_N(M) = \rho_N .$$

Since this inequality holds for all positive integers N,

$$\rho(M) = \lim_{N \to \infty} \rho_N(M) \leq \lim_{N \to \infty} \rho_N = \bar{\rho} .$$

But this is true for all strategies M, and hence

$$\rho \leq \bar{\rho} .$$

On the other hand, the inequality

$$\lim_{N \to \infty} \rho_N(M) = \rho(M) \leq \rho$$

holds for all strategies M, and so

$$\lim_{N \to \infty} \rho_N \leq \rho .$$

Therefore $\bar{\rho} \leq \rho$.

Corollary 3.12 In order to estimate ρ, it is enough to estimate ρ_N , and allow N to increase indefinitely.

One reason for our choice of exponent, is that it gives in general a good estimate of the ratio $d_{N+1}(M) / d_N(M)$, i.e., the average decrease in diameter per evaluation of the function. More precisely:

Lemma 3.13 $$\rho(M) = \lim_{N \to \infty} d_N(M) / d_{N+1}(M)$$

provided the right hand side exists.

Proof By a standard result of classical analysis, that for any sequence $d_N(M)$ of positive numbers

$$\underline{\lim}_{N \to \infty} d_{N+1}(M)/\, d_N(M) \leq \underline{\lim}_{N \to \infty} [d_N(M)]^{\frac{1}{N}} \leq \overline{\lim} [d_N(M)]^{\frac{1}{N}}$$

$$\leq \overline{\lim}_{N \to \infty} d_{N+1}(M)/d_N(M)$$

where $\underline{\lim}$ and $\overline{\lim}$ denote lower limit and upper limit respectively, it follows that

$$\lim_{N \to \infty} d_{N+1}(M) \,/\, d_N(M) = \lim_{N \to \infty} [d_N(M)]^{\frac{1}{N}}$$

$$= \lim_{N \to \infty} d_o^{\frac{1}{N}} \,/\, \rho_N(M)$$

$$= 1 \,/\, \lim_{N \to \infty} \rho_N(M)$$

$$= 1 \,/\, \rho(M)$$

Example 3.14 We consider and compare the two strategies given in section 1-3 for localizing the maximum point of a unimodal function. Take $D = [0,1] = K_o$ so that $d_o = 1$.

If N is any given integer, then $d_N = \inf_M [d_N(M)] = 1/F_N$.

So $\rho = \bar{\rho} = \lim_{N \to \infty} (F_N)^{\frac{1}{N}} = (1 + \sqrt{5})/2$. It should be noted that there is no strategy M for which $\rho_N(M) = \rho_N$, although for every $\epsilon > 0$, there is a strategy M for which $\rho_N - \epsilon \leq \rho_N(M) < \rho_N$.

If M' denotes the golden-mean strategy, then

$$d_o = d_1(M') = 1 \quad \text{and}$$

$$d_N(M') = [(\sqrt{5} - 1)/2]^{N-1} \quad \text{if} \quad N \geq 2$$

So $$\rho_N(M') = [(\sqrt{5} + 1)/ 2]^{1 - \frac{1}{N}} \quad \text{if } N \geq 2$$

and $$\rho(M') = (\sqrt{5} + 1)/ 2 = \rho = \bar{\rho} .$$

M' is an optimal strategy according to definition 3.8.

In order to compare the two strategies, let us suppose that N is given, and that functions $f \in F$ may be evaluated at just N points in D. There is the choice of either using the appropriate Fibonnacci numbers, or using the golden-mean strategy. One can compute the ratio $d_N(M')/ d_N$ or $\rho_N(M')/\rho_N$, to give some idea of how much slower the golden-mean strategy is, when employed for the case where N is given. Indeed, since

$$F_N = \frac{1}{\sqrt{5}} \left(\frac{1 + \sqrt{5}}{2} \right)^{N+1} - \frac{1}{\sqrt{5}} \left(\frac{1 - \sqrt{5}}{2} \right)^{N+1}, \text{ for } N \geq 0$$

$$d_N(M')/d_N = [\rho_N/\rho_N(M')]^N$$

$$= F_N/ \left(\frac{\sqrt{5} + 1}{2}\right)^{N-1} \quad \text{if } N \geq 2$$

$$= (3 + \sqrt{5})/(2\sqrt{5}) + [(\sqrt{5} - 3)/ 2]^N/ \sqrt{5}$$

For large N, $d_N(M') \sim d_N(1.17...)$

and $$\rho_N \sim \rho_N(M') \sqrt[N]{(1.17...)}$$

One can ask, in general, the following question. Suppose M is a strategy for which $\rho(M) = \rho = \lim_{N \to \infty} \rho_N$. For any given positive integer N, how does $\rho_N(M)$ compare with ρ_N ? In other words, if M is an optimal strategy, and M is used only until the function $f \in F$ is evaluated at N points in D, how does this procedure compare with the best procedure possible when N is given ? This appears to be a difficult question to

answer, though we present below a solution for a special case. First, a lemma is established which gives information on the behaviour of ρ_N .

Lemma 3.15 Suppose the domain D, and class F are given, such that

(i) $D = K_o$ is a compact interval on the real line with length d_o.

(ii) If K is any compact sub-interval of D, if $f \in F$, $S(f) \in K$, and X is any abscissa-set in K, and if

$K' = \{x \in K \mid x = s(g)$ for some $g \in F$ such that $f(x) = g(x)\}$

then K' is a sub-interval of K, and the ratio of diameters $d[K']/d[K]$ is invariant under all changes of coordinates of D which take the form of translation followed by a change of scale.

Then (a) $\rho_N \leq \rho$ for all positive integers N.

(b) $\rho_N \leq \rho_{mN}$ for any N and all positive integers m.

Proof It is immediate from condition (ii) that if, for some strategy M, and any k, $d_k(M) \leq d_o$, then for all positive integers m, $d_{mk}/d_o \leq [d_k(M)/d_o]^m$. It follows that

$$\rho_k(M) = [d_o/d_k]^{\frac{1}{k}} \leq [d_o/d_{mk}]^{\frac{1}{mk}} = \rho_{mk}$$

and hence $\rho_k \leq \rho_{mk}$ for all positive integers m and k. This proves (b).

To prove (a), assume that $\rho_k > \rho$ for some fixed k. Then by (b), $\rho_{mk} \geq \rho_k > \rho$ for all positive integers m. If N is any sufficiently large positive intever, then there is an m such that $mk \leq N \leq mk + k$. Since the sequence d_N is

non-increasing in N,

$$\rho_N = [d_o / d_N]^{\frac{1}{N}} \geq [d_o / d_{nk}]^{\frac{1}{N}}$$
$$\geq [\rho_{mk}]^{\frac{mk}{N}}$$
$$\geq \rho_k^{\frac{N-k}{N}} .$$

But this last inequality is impossible, since it implies that

$$\rho = \lim_{N \to \infty} \rho_N \geq \rho_k > \rho .$$

Definition 3.16 A strategy M is called periodic if for two fixed integers p and k, and some constant λ,

$$d_{p+km}(M) = \lambda^{-mk} d_p(M)$$

for all positive integers m.

Periodic strategies often arise in cases where given the domain D, and an abscissa set X in D, one gets a new domain D' and an abscissa set X' related to the former by a translation and change of scale, by evaluating any function $f \in F$ at exactly k points in D different from X. An example is the golden-mean strategy M, with p = 1, k = 1, $d_1(M) = 1$ and $\lambda = \frac{\sqrt{5} + 1}{2}$. Here, one always has an interval, with an abscissa set which consists of the two points which divide the interval in the golden ratio.

Lemma 3.17 Suppose the conditions of lemma 3.14 are satisfied, and there is an optimal periodic strategy M (which implies that $\lambda = \rho$). Then

$$1 \leq \rho_{p+mk} / \rho_{p+mk}(M) \leq c^{\frac{1}{m}}$$

for all positive integers m where c is a fixed constant $c > 1$. In other words, the difference $\rho_{p+mk} - \rho_{p+mk}(M) = 0\ (\frac{1}{m})$.

Proof

$$[\rho_{p+mk}(M)]^{p+km} = [d_o / d_{p+km}(M)]$$

$$= [d_o / d_p(M)]\rho^{mk}$$

$$= [\rho_p(M)]^p \rho^{mk}$$

Hence $$\rho_{p+km}(M) / \rho = [\rho_p(M) / \rho]^{\frac{p}{p+km}}$$

and therefore

$$1 \leq \frac{\rho_{p+mk}}{\rho_{p+km}(M)} \leq \frac{\rho}{\rho_{p+km}(M)} = \left[\frac{\rho}{\rho_p(M)}\right]^{\frac{p}{p+km}} \leq c^{\frac{1}{m}}$$

$$\text{if} \quad c = [\rho / \rho_p(M)]^{\frac{p}{k}} .$$

3 - 4 Optimal Strategies

For the final part of this chapter, we consider the question of the existence of optimal strategies. We have seen already that if the number N of observations allowed is given a priori (N finite), then there may not exist an optimal strategy. For example, consider N = 2 in the unimodal problem as in theorem 1.1, where if $d_o = 1$, then $d_2 = \frac{1}{2}$, yet $d_2(M) > \frac{1}{2}$ for every strategy M. Nevertheless, there do exist two numbers x_1 and x_2, namely $x_1 = x_2 = \frac{1}{2}$, such that the best attainable results may be obtained by choosing an abscissa set, each element of which is arbitrarily close to x_1 and x_2 respectively.

It is the purpose of this section to show that under certain conditions, there exists a point $\bar{X}$ in a Euclidean space, such that the best possible results may be obtained by selecting a point Y either equal to $\bar{X}$ or arbitrarily close to it. This point Y effectively defines the strategy M.

Definition 3.18 Let D be a given domain. For two positive integers k, and N, let

$$X = \{x_{11}; x_{21}, \ldots, x_{2k}; x_{31}, \ldots, x_{3k^2}; x_{41}, \ldots, x_{4k^3}; x_{51} \ldots; \ldots; x_{N1}, \ldots x_{Nk^{N-1}}\}$$

be an ordered set of points in D.

X can be regarded as a vector in the Cartesian product set

$$\tilde{D} = D \times D^k \times D^{k^2} \times \ldots \times D^{k^{N-1}} .$$

Define a relation $\prec$ on the elements x_{ij}, $1 \leq j \leq k^{i-1}$, of X by $x_{ij} \prec x_{i+1,j'}$, if and only if $(j-1)k < j' \leq jk$ for all $i = 1, 2, 3, \ldots (N-1)$.

Definition 3.19 Let X be as in definition 3.18 . A chain at X is an ordered sequence $x_{1j_1}, x_{2j_2}, \ldots x_{N,j_N}$ of elements of X, where

$$x_{i,j_i} \prec x_{i+1,j_{i+1}} \quad \text{for } i = 1, 2, 3, \ldots (N-1).$$

Thus, for each X, there are k^{N-1} chains, each consisting of N elements of X.

Definition 3.20 Suppose D and a class of functions F with domain D are given. Suppose X is as in definition 3.18, such

that the relation $\underset{X}{\sim}$ (see definition 2.11) partitions F at X into a finite number p of equivalence classes

$F_1(X)$, $F_2(X)$, ... $F_p(X)$.

Put $d_N[F_i(X)] = \inf_Y$ diam.$\{x \in H_o \mid x = S(g)$ for some $g \in F$ such that $f(Y) = g(Y)$, where $f \in F_i(X)$, $g \in F_i(X)$, and Y is a chain at X$\}$,

and put

$$d_N(X) = \sup_{i=1}^{p} d_N[F_i(X)] .$$

Roughly, this means that given X, $d_N(X)$ is computed by selecting for each equivalence class at X, the 'best' chain at X, and then the 'worst' case taken over all equivalence classes.

Theorem 3.21 Suppose the class F, and the domain D are given, D being a convex compact subset in the Euclidean space E^q. Assume that for every abscissa set X consisting of a finite number p of points in D, the relation $\underset{X}{\sim}$ partitions F at X into at most k^p equivalence classes, where k is a fixed positive integer.

If the function $\phi(X)$, defined by

(a) $\phi(X) = d_N(X)$, if every chain at X consists of N distinct points in D,

(b) $\phi(X) = \lim_{Y \to X} \phi(Y)$, if some chain at X consists of N-1 or fewer distinct points in D,

is a continuous real-valued function on $\tilde{D}$ ($\tilde{D}$ with metric defined by the appropriate Euclidean space), then either

(i) there is a strategy M for which $\rho_N = \rho_N(M)$

or (ii) there is a strategy M_ϵ for every $\epsilon > 0$, such that

$\rho_N(M) > \rho_N - \epsilon$, where M_ϵ is defined by points in a neighbourhood of a fixed point in $\tilde{D}$.

Proof It follows from the hypotheses of the theorem, and the preceeding definitions, that if M is a strategy for which $K_o = H_o$, every K_n is chosen as the corresponding H_n, and which involves the evaluation of functions $f \in F$ at N points of D, can be identified with a point X in $\tilde{D}$. For, one can select first a point $x_{11} \in D$, evaluate $f(x_{11})$, select the appropriate equivalence class determined by $f(x_{11})$, then the appropriate x_{2i_2} etc. Moreover, under this identification, the two numbers $d_N(M)$ of definition 3.1, and $d_N(X)$ of definition 3.20, are identical.

Since D is compact, so is $\tilde{D}$ by Tychonoff's theorem on the product of compact sets. So, since ϕ is continuous on $\tilde{D}$, ϕ attains its minimum value at some point $\bar{X}$ in $\tilde{D}$. There are two possibilities:

(i) Every chain at $\bar{X}$ consists of N distinct points in D. Then $\bar{X}$ defines a strategy M for which $\rho_N(M)$ is minimal, i.e., $\rho_N(M) = \rho_N$.

(ii) Some chain at $\bar{X}$ consists of less than N distinct points of D. In this case, for $\epsilon > 0$ and arbitrary, there is a point Y with distance ϵ of $\bar{X}$ such that every chain at Y consists of N distinct points of D. By continuity of the function $\phi(X)$, the point Y defines a strategy M for which $\rho_N(M)$ is arbitrarily close to ρ_N.

Remark For the bisection and unimodal problems, and for those presented later in chapter 5, the function ϕ in theorem 3.21 is the supremum of finitely many continuous functions, and so is continuous. Thus theorem 3.21 can be applied to these problems.

CHAPTER 4

RANDOM LOCALIZABILITY

4 - 1 Definitions

So far, we have considered ways of localizing a functional by means of an explicitly given strategy. That is, a set K_n, and possibly some abscissa sets X_i and real numbers $f(X_i)$, $f \in F$, $i = 0, 1, 2. \ldots (n-1)$ uniquely determine the next abscissa set X_n. Sometimes the uniqueness requirement can be relaxed, as in the strategy used by Gleason in theorem 1.13, where the first abscissa set X_0 need be chosen only so that discs of a certain radius with centres at the points of X_0 cover the unit square.

One asks if it is possible to localize a functional by choosing abscissa sets at random in some sense. This question is important, since optimal strategies may be too difficult to determine, or too costly to use. The more inefficient procedure of leaving the choice of abscissa sets to chance may be sufficiently satisfactory.

First, we define random localizability, preserving the terminology and notation of the previous chapters as far as possible. Our definition is essentially a generalization of the ideas given at the end of section 1-2.

We assume that D is a compact convex body in Euclidean q-dimensional space E^q. Let F be a given family of real-valued functions defined on D, such that to each function $f \in F$, there corresponds a unique point S(f) in D. Let $\phi(x)$ be a density function defined on D. That is, $\phi(x)$ is a non-negative Lebesgue integrable function which vanishes outside of D, such that $\int_D \phi(x)dx = 1$.

In the following, the terms k-abscissa set $X = \{x_1, x_2, \ldots x_k\}$, and the notation f(X), D^k etc., have the same meaning as in chapter 2.

Definition 4.1 x_o is a random variable in D distributed with density function ϕ if for each measurable subset E of D, the probability that the value of x_o belongs to E is $\int_E \phi(x)dx$.

Definition 4.2 Suppose $f \in F$. A kth order random search for S(f) is defined by the following procedure:

(i) Let $X_o = \{x_{o1}, x_{o2}, \ldots, x_{ok}\}$ be a k-abscissa set in D, the elements x_{oi} of which are independently distributed random variables, each distributed with density function ϕ.

(ii) Put $H_1(f,X_o) = \{x \in D \mid x = S(g)$ for some $g \in F$ such that $f(X_o) = g(X_o)\}$

(iii) Let $K_1(f,X_o)$ be a subset of D such that $H_1(f,X_o) \subset K_1(f,X_o)$.

(iv) Let $\phi_1(x)$ be a density function defined on $K_1(f,X_o)$.

In general, given the set $K_n(f,X_o,X_1, \ldots X_{n-1})$, and a density function $\phi_n(x)$ defined on $K_n(f,X_o,X_1, \ldots X_{n-1})$, the procedure is continued for all positive integers n, by

(i) letting $X_n = \{x_{n1}, \ldots x_{nk}\}$ be a k-abscissa set in $K_n(f,X_o, \ldots, X_{n-1})$, the elements x_{ni} of which are independent random variables on $K_n(f, X_o, X_1, \ldots X_{n-1})$, each distributed with density ϕ_n;

(ii) putting $H_{n+1}(f,X_o,X_1,\ldots X_n)=\{x\in K_n(f,X_o,X_1,\ldots X_{n-1} | x=S(g)$ for some $g\in F$ such that $f(X_n) = g(X_n)\}$;

(iii) letting $K_{n+1}(f,X_o,X_1,\ldots X_n)$ be a subset of $K_n(f,X_o,X_1,\ldots X_{n-1})$ such that

$$H_{n+1}(f,X_o,X_1, \ldots X_n) \subset K_{n+1}(f,X_o,X_1,\ldots X_n);$$

(iv) letting $\phi_{n+1}(x)$ be a density function on $K_{n+1}(f,X_o,X_1, \ldots X_n)$.

<u>Definition 4.3</u> A random strategy (usually denoted by M), is a rule, given in advance of any computations, which determines the density function ϕ, the integer k, the set K_n in terms of H_n, and the density function $\phi_n(x)$ in terms of both K_n and the density function $\phi(x)$. K_n, or $K_n(f,M)$, is an abbreviation for $K_n(f,X_o,X_1,\ldots X_{n-1})$.

If X_n is the abscissa set $X_n=\{x_{n1},x_{n2},\ldots x_{nk}\}$ and ϕ_n is the density function on K_n, we write

$$\phi_n(X_n) = \phi_n(x_{n1})\phi_n(x_{n2}) \ldots \phi_n(x_{nk})$$

and $$dX_n = dx_{n1}dx_{n2}dx_{n3} \ldots dx_{nk} .$$

With this notation, we can write an expression for the expected value of the diameter of the set $K_n(f,M)$:

$$E[K_n(f,M)] = \int_{D^k} \int_{K_1^k} \int_{K_2^k} \cdots \int_{K_{n-1}^k} d[K_n(f,X_o,X_1, \ldots X_{n-1})]$$
$$\phi_{n-1}(X_{n-1}) \ldots \phi_1(X_1)\phi(X_o)dX_{n-1}dX_{n-2}\ldots dX_o,$$

where $d[K_n(f,X_o,X_1,\ldots X_{n-1})]$ is the diameter of the set $K_n(f,X_o,X_1,\ \ldots\ X_{n-1})$.

Definition 4.4 The functional is randomly localizable with respect to the random strategy M if

$$\lim_{n \to \infty} \sup_{f \in F} E[K_n(f,M)] = 0.$$

Remark Some limitations on the scope of the previous definitions arise because the random strategy must be stated a priori. Perhaps the most natural random strategy, is that which requires each set K_n to be the smallest superset of the corresponding set H_n, which is similar to D. Then the density function $\phi_n(x)$ on K_n may be obtained by adjusting the function $\phi(x)$ by a translation and change of scale of its coordinates, with the restriction that $\int_{K_n} \phi_n(x)dx = 1$.

With the above random strategy, it is possible that some of the sets K_n may not be subsets of domain D, if D has dimension greater than 1. Occasionally, we can remove this difficulty by extending the domain D of the class of functions F, to D' ensuring that all points S(f), $f \in F$, are in D, and requiring only that $K_n \subset D'$. Or, if D is a cube, we can require not only that K_n is similar to D, but also that K_n is a subset of D.

Another possible random strategy is one which demands that every density function $\phi_n(x)$ be constant. We can then choose $K_n = H_n$ for all n.

We can modify the random search procedure as described

in definition 4.2 by allowing only l points of an abscissa set X_n to be chosen at random, where $l < k$, for $n \geq 1$. The remaining $(k-l)$ points required for X_n can be supplied by known results from previous stages. For example, if $D = [0,1]$, and if we are searching for the maximum point of a unimodal function f, then after selecting two points x_1 and x_2 in D for X_0, the abscissa set X_1 may consist of one of these points, together with just one third point x_3.

4 - 2 General theorems and examples

In this section, we consider the relationship between localizability and random localizability, and some analogues of theorems obtained in chapter 3. Intuitively, we expect that under sufficient restrictions, a functional is localizable if and only if it is randomly localizable, and that random localization is less efficient than localization.

To illustrate some of the difficulties which may arise, let us consider briefly the following problem. Let $D = [0,1]$ be the unit interval, and F the class of continuous functions f, defined on D, which are negative everywhere except for a small subinterval H_f of D on which they are positive and concave with a unique absolute maximum point at $S(f) \in H_f$.

It is easily seen that the functional $S(f)$ is not localizable, for the same reason as in the problem considered in section 2-5. The subinterval H_f over which a function $f \in F$ is positive may be too small. According to definition 4.4,

the functional is not randomly localizable either. On the other hand, localization of any point S(f) is possible after one point z at which $f(z) > 0$ is found. If H_f is always of length greater than zero, then there is a positive probability of finding such a point z, if the density function on D does not vanish anywhere.

We wish to prove now that under sufficient restrictions, localizability implies random localizability. Let us assume that all the conditions of theorem 3.3 are satisfied. Then in particular, if $S(f) \in K_o = D$, and D is a compact cube with diameter d_o, then there is an integer k, a constant c, $0 < c < 1$, and a k-abscissa set X_o, such that $H_1(f,X_o)$ can be contained in a cube $K_1(f,X_o)$ of diameter at most cd_o. This condition is not sufficient to prove that $\sup_{f\in F} E[K_1(f,M)] = \int_{D^k} d[K_1(f,M)]\phi(x)dx < d_o$. Two more are required. First, if X_o is replaced by any abscissa set Y sufficiently close to X_o, then Y is equally acceptable for localization. Secondly, the density function ϕ must be bounded away from zero near X_o. With these restrictions, we prove

Theorem 4.5 Let D and F be given, satisfying all the above conditions. Let ρ be the Euclidean metric on D^k and suppose there is a k-abscissa set X_o in D, and an $\epsilon > 0$ such that

(i) there is a constant $c < 1$, such that if Y is any k-abscissa set in $B_\epsilon = \{Y \in D^k \mid \rho(X_o, Y) < \epsilon\}$, then for all $f \in F$, $H_1(f,Y)$ can be contained in a cube $K_1(f,Y)$ where $d[K_1(f,Y)] \leq cd_o$

(ii) for some $\delta > 0$, $\delta \leq (1-c) \int_{B_\epsilon} \phi(Y)dY$.

Then $\sup_{f\in F} E[K_1(f,M)] \leq (1-\delta)d_o$.

Proof If $f \in F$, then

$$E[K_1(f,M)] = \int_{D^k} d[K_1(f,X)]\phi(X)dX$$

$$= \int_{D^k - B_\epsilon} d[K_1(f,X)]\phi(X)dX + \int_{B_\epsilon} d[K_1(f,X)]\phi dX$$

$$\leq d_o \int_{D^k - B_\epsilon} \phi(X)dX + cd_o \int_{B_\epsilon} \phi(X)dX$$

$$= d_o \left[1 - (1-c) \int_{B_\epsilon} \phi(X)dX\right]$$

$$\leq (1 - \delta)d_o$$

Hence

$$\sup_{f \in F} E[K_1(f,M)] \leq (1 - \delta)d_o.$$

Corollary 4.6 If, under the conditions of theorem 4.5, the density functions $\phi_n(x)$ are defined as indicated in the remark following definition 4.4, then

$$\sup_{f \in F} E[K_n(f,M)] \leq (1 - \delta)^n d_o$$

for all positive integers n.

Example 4.7 We show that if $D = [0,1]$ and F is the class of unimodal functions defined in section 1-3, then the functional defined by $S(f)$, $f \in F$, is randomly localizable.

Suppose $0 < x < y < 1$, and $f \in F$. Then

$$d[K_1(f,(x,y))] = \begin{cases} y & \text{if } f(x) > f(y) \\ 1-x & \text{if } f(x) < f(y) \\ y-x & \text{if } f(x) = f(y) \end{cases}.$$

If $\phi(x)$ is a density function defined on D, then

$$E[K_1(f,M)] = \int_R \int d[K_1(f,(x,y))]\phi(x)\phi(y)\, dx\, dy$$

where R is the square $\{(x,y) \mid 0 \le x,y \le 1\}$, and so

$$\sup_{f\in F} E[K_1(f,M)] \le \int_R \int \sup_{f\in F} d[K_1(f,(x,y))]\phi(x)\phi(y)\, dx\, dy$$

$$\le \int_0^1 \int_0^y [\sup\{1-x,\ y\}\ \phi(x)\phi(y)\, dx\, dy$$

$$+ \int_0^1 \int_0^x [\sup\{1-y,x\}]\ \phi(x)\phi(y)\, dy\, dx$$

$$= 2 \int_0^{1/2} \int_x^{1-x} (1-x)\phi(x)\phi(y)\, dy\, dx$$

$$+ 2 \int_{1/2}^1 \int_{1-y}^y y\ \phi(x)\ \phi(y)\, dx\, dy\ .$$

In particular, if $\phi(x) = 1$ for $0 \le x \le 1$, then $\sup_{f\in F} E[K_1(f,M)] \le \frac{5}{6}$, and, if the assumptions of theorem 4.5 and corollary 4.6 are made,

$$\sup_{f\in F} E[K_1(f,M)] \le \left(\frac{5}{6}\right)^n .$$

For comparing the efficiency of random localization with localization, we can define a random exponent, analogous to the exponent defined in definition 3.8. Suppose the domain D, and class F are given, along with a random strategy M. For each positive integer n, put

$$\rho_{kn}(M) = [d_o / \sup_{f\in F} E[K_n(f,M)]\ ]^{\frac{1}{kn}}$$

where d_o is the diameter of D, and k is the number of points in D belonging to each abscissa set. Define the random exponent with respect to M, to be the number

$$\rho(M) = \lim_{n \to \infty} \rho_{kn}(M) .$$

There seems to be no point in defining the number $\rho = \sup_M \rho(M)$, since usually one is concerned with a particular density function ϕ. Indeed, it is shown later than the more efficient random strategies for the bisection problem are those for which ϕ has a large and thin peak about the mid-point of D. The limiting case arises when ϕ is the Dirac δ-distribution, which yields the bisection method described in section 1-1.

4 - 3 Random search for a zero

By far the simplest problem in considering random localizability is that of searching for the unique zero of a suitable function defined in a compact interval. We shall be concerned with this particular problem for the remainder of the chapter. Three main results are obtained. First, three formulas are derived for the joint distribution of the end points of the intervals to which the zero of the function is successively confined. Secondly, the expected values of the lengths of these intervals, and other associated quantities are computed. Finally, the behaviour of these expected values for large n is investigated, and some special cases are considered.

It is assumed throughout that the domain D is the closed

interval [0,1], and that the class F defined on D consists of all real-valued continuous functions f for which $f(0) > 0 > f(1)$, and which possess a unique zero $\alpha = S(f)$ in D. It is also assumed that $\phi(x)$ is a given continuous density function defined on D. In particular, $\phi(x) \geq 0$ if $0 \leq x \leq 1$, and $\int_0^1 \phi(x)dx = 1$.

By the density function on the interval [a,b], where $0 \leq a < b \leq 1$, is meant the function $\frac{1}{b-a}\phi(\frac{x-a}{b-a})$, defined for $a \leq x \leq b$. This function is just $\phi(x)$ suitably adjusted by a translation and change of scale.

Suppose $f \in F$. The random search procedure $\alpha = S(f)$ commences by letting x be a point (abscissa set) in D, chosen at random with density function ϕ. It can always be decided whether $\alpha < x_1$, $\alpha = x_1$, or $\alpha > x_1$, by considering the sign of $f(x)$. So one can put

$$l_1 = l_1(\alpha,x_1) = 0 \text{ and } r_1 = r_1(\alpha,x_1) = x, \text{ if } x_1 > \alpha$$

$$l_1 = l_1(\alpha,x_1) = x_1 \text{ and } r_1 = r_1(\alpha,x_1) = 1, \text{ if } x_1 < \alpha$$

$$l_1 = l_1(\alpha,x_1) = x_1 \text{ and } r_1 = r_1(\alpha,x_1) = x_1, \text{ if } x_1 = \alpha .$$

In any of these cases, $\alpha \in [l_1, r_1]$.

Let x_2 be a point in this interval, chosen at random with respect to the density function ϕ_1 on $[l_1, r_1]$, and define l_2 and r_2 similarly. Successively, let x_3, x_4, ...x_n, be a sequence of points where, for all positive integers n, x_{n+1} is distributed at random with respect to the density function on $[l_n, r_n]$, and

$$L_{n+1} = L_n , \quad r_{n+1} = x_{n+1} , \quad \text{if } x_{n+1} > \alpha,$$

$$L_{n+1} = x_n, \quad r_{n+1} = r_n , \quad \text{if } x_{n+1} < \alpha,$$

$$L_{n+1} = r_{n+1} = x_{n+1} , \quad \text{if } x_{n+1} = \alpha.$$

For completeness, put $L_o = 0$, $r_o = 1$. It is obvious that for all positive integers n,

$$0 \le L_n \le L_{n+1} \le \alpha \le r_{n+1} \le r_n \le 1.$$

Both L_n and r_n are functions of α, and the sequence $x_1, x_2, \ldots, x_n$.

It is convenient at this stage to define the function $F(\xi, s, t, \eta)$ for $0 \le \xi \le s \le t \le \eta \le 1$, by the equations

$$F(\xi, s, t, \eta) = \int_0^{\frac{s-\xi}{\eta-\xi}} \phi(x)\, dx + \int_{\frac{t-\xi}{\eta-\xi}}^1 \phi(x)\, dx , \text{ if } \xi < \eta .$$

$$F(\xi, s, t, \eta) = 1 , \text{ if } \xi = \eta .$$

The usefulness of this function is evident from the following lemma.

Lemma 4.8 If $\xi = L_n < \eta = r_n$ for some n, then the probability that either $\xi \le x_{n+1} < s$ or $t < x_{n+1} < \eta$ is $F(\xi, s, t, \eta)$.

Proof The stated probability is

$$\int_\xi^s \frac{1}{\eta-\xi} \phi\left(\frac{x-\xi}{\eta-\xi}\right) dx + \int_t^\eta \frac{1}{\eta-\xi} \phi\left(\frac{x-\xi}{\eta-\xi}\right) dx$$

$$= \int_0^{\frac{s-\xi}{\eta-s}} \phi(u)\, du + \int_{t-\xi}^1 \phi(u)\, du = F(\xi, s, t, \eta).$$

4 - 4 Joint probability distribution

Of basic importance is the joint probability distribution of the end points l_n and r_n of the interval $[l_n, r_n]$ which contains the zero α. This is given in terms of the function $G_n(s,t)$ below.

Definition 4.9 Let $f \in F$, and suppose α is the zero of f. Let s and t be any two numbers such that $0 \leq s \leq \alpha \leq t \leq 1$. For each integer $n \geq 0$, define $G_n(s,t)$ to be the probability that both $l_n < s$ and $t < r_n$.

Remark $G_n(0,t) = G_n(s,1) = 0$ and $G_n(\alpha,\alpha) = 1$ for all n.

We prove now the fundamental recurrence relation satisfied by $G_n(s,t)$.

Theorem 4.10

$G_o(s,t) = 1$ if both $s \neq 0$, $t \neq 0$, and

$G_o(0,t) = G_o(s,1) = 0$ for any s,t.

For each non-negative integer k,

$$G_{k+1}(s,t) = \iint F(\xi, s, t, \eta) \, dG_k(\xi, \eta)$$

where the range of integration is over the rectangle $0 \leq \xi \leq s$, $t \leq \eta \leq 1$. The sense in which the Riemann-Stieltjes integral is to be taken is described in the proof.

Proof Since $l_o(\alpha) = 0$ and $r_o(\alpha) = 1$, the statements concerning $G_o(s,t)$ are obvious.

For arbitrary $\epsilon > 0$, take partitions

$$0 = y_o < y_1 < y_2 < y_3 < \ldots < y_n = s$$

$$t = z_o < z_1 < z_2 < z_3 < \ldots < z_m = 1$$

of the intervals [0,s] and [t,1] respectively, such that

$$\sup_{i=1}^{n} \; | \, y_i - y_{i-1} \, | \; < \; \epsilon$$

$$\sup_{j=1}^{m} \; | \, z_i - z_{i-1} \, | \; < \; \epsilon \; .$$

By mutual exclusion, the probability that both $y_{i-1} \leq l_k < y_i$ and $z_{j-1} < r_k \leq z_j$ is

$$G_k(y_i, z_{j-1}) - G_k(y_i, z_j) - G_k(y_{i-1}, z_{j-1}) + G_k(y_{i-1}, z_j)$$

where $1 \leq i \leq n$, $1 \leq j \leq m$.

Moreover, if it is given that both

$$l_k < s \quad \text{and} \quad t < r_k \; ,$$

then the probability that both

$$l_{k+1} < s \quad \text{and} \quad t < r_{k+1}$$

is just the probability that either

$$l_k \leq x_{k+1} < s \quad \text{or} \quad t < x_{k+1} \leq r_k .$$

By lemma 4.8, this equals $F(l_k, s, t, r_k)$.

It follows, by adding over all pairs of intervals in the partitions and letting $\epsilon \to 0$, that

$$G_{k+1}(s,t) = \lim_{\epsilon \to 0} \sum_{i=1}^{n} \sum_{j=1}^{m} F(\xi_i, s, t, \eta_j)$$

$$[G_k(y_i, z_{j-1}) - G_k(y_i, z_j) - G_k(y_{i-1}, z_{j-1} + G_k(y_{j-1}, z_j)]$$

where for each i and j, ξ_i and η_j are arbitrary points such that

$$y_{i-1} \leq \xi_i < y_i$$

$$z_{j-1} < \eta_j \leq y_j \; .$$

The limit is taken over all possible partitions as $\epsilon \to 0$. This defines to double integral in the statement of the theorem.

It remains to show that the above limit exists. It is enough, by a standard theorem on Riemann-Stieltjes Integrals on the plane, see for example, Hildebrandt [4], to show that $F(\xi, s, t, \eta)$ is continuous in ξ and η and that $G_k(y,z)$ is of bounded variation on R.

Since $\phi(x)$ is continuous, it is immediate that $F(\xi, s, t, \eta)$ is continuous for fixed s,t.

The variation of G_k is

$$\lim_{\epsilon \to 0} \sum_{i=1}^{n} \sum_{j=1}^{m} \left| [G_k(y_i, z_{j-1}) - G_k(y_i, z_j) - G_k(y_{i-1}, z_{j-1}) + G_k(y_{i-1}, z_j)\right|$$

$$= \lim_{\epsilon \to 0} \sum_{i=1}^{n} \sum_{j=1}^{m} [G_k(y_i, z_{j-1}) - G_k(y_i, z_j) - G_k(y_{i-1}, z_{j-1}) + G_k(y_{i-1}, z_j)] \ ;$$

the non-negativity resulting from the fact that the expression in square brackets represents the probability of an event.

So the variation of G_k

$$= \lim_{\epsilon \to 0} [G_k(y_n, z_o) - G_k(y_n, z_m) - G_k(y_o, z_o) + G_k(y_o, z_m)]$$

$$= G_k(s,t) - G_k(s,1) - G_k(0,t) + G_k(0,1) = G_k(s,t) \leq 1.$$

It follows that G_k is of bounded variation, and hence the integral exists for each k.

The recurrence relation of theorem 4.10 is too difficult to use for developing properties of $G_k(s,t)$ and other functions

which will concern us later. We aim to prove two other recurrence relations for $G_k(s,t)$, the first of which is obtained from theorem 4.10 by an integration by parts. First we require a lemma. The proof is omitted.

Lemma 4.11

$$\frac{\partial F(\xi,\ s,\ t,\ \eta\ )}{\partial \xi} \quad \text{and} \quad \frac{\partial F(\xi, s,\ t, \eta\ )}{\partial \eta}$$

are continuous functions of ξ and η in the rectangle $0 \leq \xi \leq s$, $t \leq \eta \leq 1$.

In particular

$$\frac{\partial F(s,\ s,\ t, \eta\ )}{\partial \eta} = \frac{(t-s)}{(\eta - s)^2}\ \phi\left(\frac{t-s}{\eta-s}\right)$$

$$\frac{\partial F(\xi,\ s,\ t,\ t)}{\partial \xi} = -\frac{(t-s)}{(t-\xi)^2}\ \phi\left(\frac{s-\xi}{t-\xi}\right) .$$

For brevity in what follows, we write

$F(\xi, \eta\ )$ in place of $F(\xi,\ s,\ t, \eta\ )$ and

$\dfrac{\partial F(\xi,\eta\ )}{\partial \xi}$ in place of $\dfrac{\partial F(\xi,\ s,\ t, \eta\ )}{\partial \xi}$.

Theorem 4.12 If k is a non-negative integer, then

$$G_{k+1}(s,t) = -\int_0^s G_k(\xi,t)\ \frac{\partial F(\xi,t)}{\partial \xi}\ d\xi + \int_t^1 G_k(s,\eta\ )\ \frac{\partial F(s,\eta\ )}{\partial \eta}\ d\eta + \iint_R G_k(\xi,\ \eta\ )\ dF(\xi,\ \eta\ )$$

where R is the rectangle $0 \leq \xi \leq s$, $t \leq \eta \leq 1$.

The double integral is defined as the limit of the Riemann sum as $\epsilon \to 0$ of

$$\sum_{i=1}^{n-1}\sum_{j=1}^{m-1} G_k(y_i,z_j)[F(\xi_{i+1},\eta_j)-F(\xi_i,\eta_j)-F(\xi_{i+1},\eta_{j+1}) + F(\xi_i,\eta_{j+1})]$$

where

$$0 = y_0 \le \xi_1 \le y_1 \le \xi_2 \le \cdots \le \xi_n \le y_n = s$$

$$t = z_0 \le \eta_1 \le z_1 \le \eta_2 \le \cdots \le \eta_m \le z_m = 1$$

and

$$\sup_{i=1}^{n} |y_i - y_{i-1}| < \epsilon , \quad \sup_{j=1}^{m} |z_j - z_{j-1}| < \epsilon$$

Proof The approximating Riemann sum for $\iint F(\xi,s,t,\eta)dG_k(\xi,\eta)$ in theorem 4.10 may be rearranged to give

$$\sum_{i=1}^{n-1}\sum_{j=1}^{m-1} [F(\xi_i,\eta_{j+1}) - F(\xi_i,\eta_j) - F(\xi_{i+1},\eta_{j+1}) + F(\xi_{i+1},\eta_j)]G_k(y_i,z_j)$$

$$+ \sum_{j=1}^{m-1} [F(\xi_n,\eta_{j+1}) - F(\xi_n,\eta_j)]G_k(y_n,z_j)$$

$$- \sum_{j=1}^{m-1} [F(\xi_1,\eta_{j+1}) - F(\xi_1,\eta_j)]G_k(y_0,z_j)$$

$$+ \sum_{i=1}^{n} F(\xi_i,\eta_0)[G_k(y_i,z_0) - G_k(y_{i-1},z_0)]$$

$$- \sum_{i=1}^{n} F(\xi_i,\eta_m)[G_k(y_i,z_m) - G_k(y_{i-1},z_m)] .$$

Using the facts that $y_0 = 0$, $y_n = s$, $z_0 = t$, $y_m = 1$, and $G_k(0,t) = G_k(s,1) = 0$, the sum simplifies to

$$\sum_{i=1}^{n-1}\sum_{j=1}^{m-1}[F(\xi_i,\eta_{j+1})-F(\xi_i,\eta_j)-F(\xi_{i+1},\eta_{j+1})$$

$$+F(\xi_{i+1},\eta_j)]\,G_k(y_i,z_j)$$

$$+\sum_{j=1}^{m-1}[F(\xi_n,\eta_{j+1})-F(\xi_n,\eta_j)]G_k(s,z_j)$$

$$+\sum_{i=1}^{n-1}[F(\xi_i,\eta_1)-F(\xi_{i+1},\eta_1)]G_k(y_i,t)$$

$$+F(\xi_n,\eta_1)G_k(s,t).$$

Now let the mesh ϵ of this sum approach zero. Since F is continuous

$$\lim_{\epsilon\to 0} F(\xi_n,\eta_1)G_k(s,t) = F(s,t)G_k(s,t)$$

$$= F(s,s,t,t)G_k(s,t)$$

$$= 0\,.$$

By the mean-value theorem of differential calculus,

$$\sum_{j=1}^{m-1}[F(\xi_n,\eta_{j+1})-F(\xi_n,\eta_j)]G_k(s,z_j)$$

$$=\sum_{j=1}^{m-1}(\eta_{j+1}-\eta_j)\,\frac{\partial F(\xi_n,\phi_j)}{\partial\eta}\,G_k(s,z_j)\quad\text{where }\eta_j<\phi_j<\eta_{j+1}\,.$$

The difference between this expression and

$$\sum_{j=1}^{m-1}(\eta_{j+1}-\eta_j)\,\frac{\partial F(s,z_j)}{\partial\eta}\,G_k(s,z_j),$$

which is the approximating Riemann sum to

$$\int_t^1 \frac{\partial F(s,\eta)}{\partial\eta}\,G_k(s,\eta)\,d\eta\,,$$

is less than or equal to

$$\sum_{j=1}^{m-1} \left| \frac{\partial F(s,z_j)}{\partial \eta} - \frac{\partial F(\xi_n,\phi_j)}{\partial \eta} \right| G_k(s,z_j)(\eta_{j+1}-\eta_j)$$

which approaches 0 as $\epsilon \to 0$ by lemma 4.13.

So

$$\lim_{\epsilon \to 0} \sum_{j=1}^{m-1} [F(\xi_n,\eta_{j+1}) - F(\xi_n,\eta_j)]G_k(s,z_j) = \int_t^1 G_k(s,\eta)\frac{\partial F(s,\eta)}{\partial \eta}d\eta$$

Similarly

$$\lim_{\epsilon \to 0} \sum_{i=1}^{n-1} [F(\xi_i,\eta_o) - F(\xi_{i+1},\eta_o)]G_k(y_i,t)$$

$$= -\int_o^s \frac{\partial F(\xi,t)}{\partial \xi} G_k(\xi,t)d\xi.$$

Finally, the sum of all the remaining terms must approach a limit, and this sum is what is defined to be

$$\iint_R G_k(\xi,\eta)dF(\xi,\eta).$$

We can now derive a third recurrence relation for $G_k(s,t)$. It is the simplest one to use in this chapter. We claim that for all integers $k \geq 0$,

$$G_{k+1}(s,t) = \int_0^s G_k\left(\frac{s-x}{1-x}, \frac{t-x}{1-x}\right) \phi(x)dx + \int_t^1 G_k\left(\frac{s}{x}, \frac{t}{x}\right) \phi(x)dx.$$

It does not seem possible to prove this result directly from theorem 4.12, so we obtain it by induction on k. It is convenient to introduce some new notation, and rephrase the result of theorem 4.12. For each integer $k \geq 0$, we put

$$A_k(s,t) = \int_t^1 \frac{G_k(s,\eta)}{(\eta-s)^2} \phi\left(\frac{t-s}{\eta-s}\right) d\eta$$

$$B_k(s,t) = \int_0^s \frac{G_k(\xi,t)}{(t-\xi)^2} \phi\left(\frac{s-\xi}{t-\xi}\right) d\xi$$

$$C_k(s,t) = \iint_R G_k(\xi,\eta)\, dF(\xi,\eta) \quad .$$

By lemma 4.11, we can rewrite the result of theorem 4.12 by the equation

$$G_{k+1}(s,t) = (t-s)[A_k(s,t) + B_k(s,t)] + C_k(s,t)$$

Theorem 4.13

If the functions $H_k(s,t)$, $k = 0, 1, 2, \ldots$ are defined on the rectangle $0 \le s \le \alpha$, $\alpha \le t \le 1$ by the equations

$$H_o(0,t) = H_o(s,1) = 0$$

$$H_o(s,t) = 1, \text{ if } 0 < s \le t < 1$$

$$H_{k+1}(s,t) = \int_0^s H_k\left(\frac{s-x}{1-x}, \frac{t-x}{1-x}\right) \phi(x)dx + \int_t^1 H_k\left(\frac{s}{x}, \frac{t}{x}\right) \phi(x)dx,$$

then $H_k(s,t) = G_k(s,t)$, if $0 \le s \le \alpha$ and $\alpha \le t \le 1$, for each positive integer .

Proof The theorem holds for $k = 0$, by the definitions. If $k = 1$, then

$$H_1(s,t) = \int_0^s \phi(x)dx + \int_t^1 \phi(x)dx = F(0,s,t,1).$$

From theorem 4.10,

$$G_1(s,t) = \lim_{\epsilon \to 0} \sum_{i=1}^{n} \sum_{j=1}^{m} F(\xi_i, s, t, \eta_j)[G_o(\eta_i, z_{j-1}) - G_o(\eta_i, z_j)$$

$$- G_o(\eta_{i-1}, z_{j-1}) + G_o(\eta_{j-1}, z_j)]$$

$$= \lim_{\epsilon \to 0} F(\xi_1, s, t, \eta_m) G_o(x_1, \eta_{m-1})$$

all other terms of the sum being zero, since

$$G_o(s,t) = 1 \quad \text{if} \quad 0 < s \leq t < 1 .$$

So

$$G_1(s,t) = \lim_{\epsilon \to 0} F(\xi_1, s, t, \eta_m)$$

$$= F(0,s,t,1) = H_1(s,t).$$

Assume the induction hypothesis that for some integer $k \geq 0$,

$$H_k(s,t) = G_k(s,t)$$

and

$$H_{k+1}(s,t) = G_{k+1}(s,t).$$

We prove under this assumption first that

$$C_{k+1}(s,t) = \int_o^s C_k\left(\frac{s-x}{1-x}, \frac{t-x}{1-x}\right) \phi(x)dx + \int_t^1 C_k\left(\frac{s}{x}, \frac{t}{x}\right) \phi(x)dx,$$

and then that

$$H_{k+2}(s,t) = G_{k+2}(s,t) .$$

By definition of C_{k+1}, and the induction hypothesis,

$$C_{k+1}(s,t) = \iint_R G_{k+1}(\xi, \eta)dF(\xi, \eta)$$

$$\iint_R H_{k+1}(\xi, \eta)dF(\xi, \eta)$$

$$= \int_{\eta=t}^{1} \int_{\xi=0}^{s} \int_{0}^{\xi} H_k\left(\frac{\xi-x}{1-x}, \frac{\eta-x}{1-x}\right) \phi(x)\,dx\,dF(\xi,\eta)$$

$$+ \int_{\eta=t}^{1} \int_{\xi=0}^{s} \int_{\xi}^{1} H_k\left(\frac{s}{x}, \frac{\eta}{x}\right) \phi(x)\,dx\,dF(\xi,\eta)$$

$$= \int_{0}^{s} \int_{\eta=t}^{1} \int_{\xi=x}^{s} H_k\left(\frac{\xi-x}{1-x} \frac{\eta-x}{1-x}\right) \phi(x)\,dF(\xi,\eta)\,dx$$

$$+ \int_{t}^{1} \int_{\xi=0}^{s} \int_{\eta=t}^{x} H_k\left(\frac{\xi}{x}, \frac{\eta}{x}\right) \phi(x)\,dF(\xi,\eta)\,dx$$

by changing the order of integration. So

$$C_{k+1}(s,t) = \int_{0}^{s} \int_{v=\frac{t-x}{1-x}}^{1} \int_{v=0}^{\frac{s-x}{1-x}} H_k(u,v)\,dF(u,v)\ \phi(x)\,dx$$

$$+ \int_{t}^{1} \int_{v=0}^{\frac{s}{x}} \int_{v=\frac{t}{x}}^{1} H_k(u,v)\,dF(u,v)\,\phi(x)\,dx$$

$$= \int_{0}^{s} C_k\left(\frac{s-x}{1-x}, \frac{t-x}{1-x}\right) \phi(x)\,dx$$

$$+ \int_{t}^{1} C_k\left(\frac{s}{x}, \frac{t}{x}\right) \phi(x)\,dx$$

since $H_k = G_k$ by hypothesis.

To complete the proof, we have

$$G_{k+2}(s,t) - C_{k+1}(s,t)$$

$$= (t-s)A_{k+1}(s,t) + (t-s)B_{k+1}(s,t)$$

$$= (t-s)\int_{t}^{1} H_{k+1}(s,\eta)\frac{1}{(\eta-s)^2}\ \phi\left(\frac{t-s}{\eta-s}\right) dy$$

$$+ (t-s)\int^{s} H_{k+1}(\xi,t)\frac{1}{(t-\xi)^2}\ \phi\left(\frac{s-\xi}{t-\xi}\right) d\xi$$

$$= (t-s) \int_t^1 \int_o^s G_k\left(\frac{s-x}{1-x}, \frac{\eta-x}{1-x}\right) \frac{1}{(\eta-s)^2} \phi\left(\frac{t-s}{\eta-s}\right) \phi(x)\,dx\,d\eta$$

$$+ (t-s) \int_t^1 \int_\eta^1 G_k\left(\frac{s}{x}, \frac{\eta}{x}\right) \frac{1}{(\eta-s)^2} \phi\left(\frac{t-s}{\eta-s}\right) \phi(x)\,dx\,d\eta$$

$$+ (t-s) \int_o^s \int_o^\xi G_k\left(\frac{\xi-x}{1-x}, \frac{t-x}{1-x}\right) \frac{1}{(t-\xi)^2} \phi\left(\frac{s-\xi}{t-\xi}\right) \phi(x)\,dx\,d\xi$$

$$+ (t-s) \int_o^s \int_t^1 G_k\left(\frac{\xi}{x}, \frac{t}{x}\right) \frac{1}{(t-\xi)^2} \phi\left(\frac{s-\xi}{t-\xi}\right) \phi(x)\,dx\,d\xi$$

by using the assumption that $H_k = G_k$, and $H_{k+1} = G_{k+1}$.

By a suitable change of variables in each of these four integrals,

$$G_{k+2}(s,t) - G_{k+1}(s,t)$$

$$= (t-s) \int_o^s \int_{\frac{t-s+u(1-t)}{1-s}}^1 \frac{1}{1-u} \phi\left(\frac{s-u}{1-u}\right) \frac{G_k(u,v)}{(v-u)^2} \phi\left[\frac{(t-s)(1-u)}{(v-u)(1-s)}\right] dv\,du$$

$$+ (t-s) \int_s^{\frac{s}{t}} \int_{\frac{ut}{s}}^1 \frac{1}{u} \phi\left(\frac{s}{u}\right) \frac{G_k(u,v)}{(v-u)^2} \phi\left[\frac{u(t-s)}{s(v-u)}\right] dv\,du$$

$$+ (t-s) \int_t^1 \int_o^{\frac{sv}{t}} \frac{1}{v} \phi\left(\frac{t}{v}\right) \frac{G_k(u,v)}{(v-u)^2} \phi\left[\frac{vs-ut}{t(v-u)}\right] du\,dv$$

$$+ (t-s) \int_{\frac{t-s}{1-s}}^t \int_0^{\frac{v(1-s)-(t-s)}{1-t}} \frac{1}{1-v} \phi\left(\frac{t-v}{1-v}\right) \frac{G_k(u,v)}{(v-u)^2} \phi\left[\frac{v(1-s)-u(1-t)-(t-s)}{(v-u)(1-t)}\right] du\,dv$$

$$= (t-s) \int_o^s \frac{1}{1-u} \phi\left(\frac{s-u}{1-u}\right) A_K\left(u, \frac{(t-s)+u(1-t)}{1-s}\right) du$$

$$+(t-s)\int_{s}^{\frac{s}{t}} \frac{1}{u}\,\phi\left(\frac{s}{u}\right) A_k\left(u, \frac{ut}{s}\right)\,du$$

$$+(t-s)\int_{\frac{t-s}{1-s}}^{t} \frac{1}{1-v}\,\phi\left(\frac{t-v}{1-v}\right) B_k\left(\frac{v(1-s)-(t-s)}{1-t}, v\right)\,dv$$

$$+(t-s)\int_{t}^{1} \frac{1}{v}\,\phi\left(\frac{t}{v}\right) B_k\left(\frac{sv}{t}, v\right)\,dv$$

$$=(t-s)\int_{o}^{s} \frac{1}{1-x}\,\phi(x)\, A_k\left(\frac{s-x}{1-x}, \frac{t-x}{1-x}\right)\,dx$$

$$+(t-s)\int_{t}^{1} \frac{1}{x}\,\phi(x)\, A_k\left(\frac{s}{x}, \frac{t}{x}\right)\,dx$$

$$+(t-s)\int_{o}^{s} \frac{1}{1-x}\,\phi(x)\, B_k\left(\frac{s-x}{1-x}, \frac{t-x}{1-x}\right)dx$$

$$+(t-s)\int_{t}^{1} \frac{1}{x}\,\phi(x)\, B_k\left(\frac{s}{x}, \frac{t}{x}\right)\,dx$$

$$=\int_{o}^{s} \phi(x)\left[G_{k+1}\left(\frac{s-x}{1-x}, \frac{t-x}{1-x}\right) - C_k\left(\frac{s-x}{1-x}, \frac{t-x}{1-x}\right)\right]\,dx$$

$$+\int_{t}^{1} \phi(x)\left[G_{k+1}\left(\frac{s}{x}, \frac{t}{x}\right) - C_k\left(\frac{s}{x}, \frac{t}{x}\right)\right]\,dx$$

$$=\int_{o}^{s} G_{k+1}\left(\frac{s-x}{1-x}, \frac{t-x}{1-x}\right)\phi(x)\,dx + \int_{t}^{1} G_{k+1}\left(\frac{s}{x}, \frac{t}{x}\right)\phi(x)\,dx$$
$$-C_{k+1}(s,t).$$

Finally, the functions $H_K(s,t) = G_K(s,t)$ are all non-negative and continuous in the rectangle $0 < s < \alpha$, $\alpha < t < 1$ by induction on K, and so all the integrands are absolutely integrable. So the changes of order of integration are valid. The mappings

$$u = \frac{s-x}{1-x} \quad , \quad v = \frac{t-x}{1-x}$$

and

$$u = \frac{s}{x} \quad , \quad v = \frac{t}{x}$$

with x regarded as a parameter, are all 1-1 with non-negative Jacobians, and so the changes of variables used above are also justified. Thus

$$G_{k+2}(s,t) = \int_0^s G_{k+1}\left(\frac{s-x}{1-x}, \frac{t-x}{1-x}\right) \phi(x)dx + \int_t^1 G_{k+1}\left(\frac{s}{x}, \frac{t}{x}\right)\phi(x)dx$$

as claimed. This completes the proof of the theorem.

4 - 5 Expected values

We are now in a position to compute the moments and expected values of various functions of l_n and r_n. In particular, we find formulas for the expected value and variance of $r_n - l_n$, the knowledge of which allows a comparison with the bisection method of localization. Since both l_n and r_n depend on α as well as on the abscissa sets $x_1, x_2, \ldots, x_n$, the expected values are functions of α. We use the symbol E for expected value. R denotes the rectangle $\{(x,y) \quad 0 \leq x \leq \alpha, \quad \alpha \leq y \leq 1 \}$.

Definition 4.14 For each non-negative integer n,

$$y_n(\alpha) = E(l_n) = \iint_R x \, dG_n(x,y)$$

$$z_n(\alpha) = E(r_n) = \iint_R y \, dG_n(x,y)$$

$$e_n(\alpha) = E(r_n - l_n) = z_n(\alpha) - y_n(\alpha)$$

$$w_n(\alpha) = E(r_n + l_n) = y_n(\alpha) + z_n(\alpha)$$

$$p_n(\alpha) = E(l_n r_n) = \iint_R xy \, dG_n(x,y)$$

$$h_n(\alpha) = E(l_n^2) = \iint_R x^2 dG_n(x,y)$$

$$k_n(\alpha) = E(r_n^2) = \iint_R y^2 \, dG_n(x,y)$$

$$f_n(\alpha) = E(r_n - l_n)^2 = h_n(\alpha) + k_n(\alpha) - 2p_n(\alpha)$$

$$v_n(\alpha) = f_n(\alpha) - [e_n(\alpha)]^2$$

$v_n(\alpha)$ is of course the variance of $r_n - l_n$.

All the above Stieltjes Integrals are to be taken in the same sense as that in theorem 4.10.

We derive recurrence relations for some of the above functions. These are later used to find asymptotic estimates for $e_n(\alpha)$ and $v_n(\alpha)$.

Theorem 4.15 $y_o(\alpha) = 0$. For each integer $n \geq 0$,

$$y_{n+1}(\alpha) = \int_o^\alpha t\phi(t)dt + \int_o^\alpha (1-t)y_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt + \int_\alpha^1 ty_n\left(\frac{\alpha}{t}\right)\phi(t)dt.$$

Proof
$$y_n(\alpha) = \iint_R x dG_n(x,y)$$
$$= \int_o^\alpha x dG_n(x,\alpha) - \int_o^\alpha x dG_n(x,1)$$
$$= \int_o^\alpha x dG_n(x,\alpha), \quad \text{since } G_n(x,1) = 0.$$

So
$$y_n(\alpha) = \alpha - \int_o^\alpha G_n(x,\alpha)dx \quad \text{on integrating by parts.}$$

Hence, from theorem 4.13,

$$\alpha - y_{n+1}(\alpha) = \int_o^\alpha G_{n+1}(x,\alpha)dx$$
$$= \int_o^\alpha \int_o^x G_n\left(\frac{x-t}{1-t}, \frac{\alpha-t}{1-t}\right)\phi(t)dtdx + \int_o^\alpha \int_\alpha^1 G_n\left(\frac{x}{t}, \frac{\alpha}{t}\right)\phi(t)dt\,dx$$
$$= \int_o^\alpha \int_t^\alpha G_n\left(\frac{x-t}{1-t}, \frac{\alpha-t}{1-t}\right)\phi(t)dx\,dt + \int_\alpha^1 \int_0^\alpha G_n\left(\frac{x}{t}, \frac{\alpha}{t}\right)\phi(t)\,dx\,dt$$
$$= \int_o^\alpha \int_0^{\frac{\alpha-t}{1-t}} (1-t)G_n\left(u, \frac{\alpha-t}{1-t}\right)\phi(t)du\,dt + \int_\alpha^1 \int_0^{\frac{\alpha}{t}} tG_n\left(u, \frac{\alpha}{t}\right)\phi(t)\,du\,dt$$

$$= \int_0^{\alpha} (1-t)\left[\frac{\alpha-t}{1-t} - y_n\left(\frac{\alpha-t}{1-t}\right)\right]\phi(t)\,dt$$

$$+ \int_{\alpha}^{1} t\left[\frac{\alpha}{t} - y_n\left(\frac{\alpha}{t}\right)\right]\phi(t)\,dt.$$

So

$$y_{n+1}(\alpha) = \alpha - \int_0^{\alpha}(\alpha-t)\phi(t)dt - \int_{\alpha}^{1}\alpha\phi(t)dt$$

$$+ \int_0^{\alpha}(1-t)y_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt + \int_{\alpha}^{1} t\, y_n\left(\frac{\alpha}{t}\right)\phi(t)dt\,,$$

and since $\int_0^1 \phi(t)dt = 1$, the result follows.

In exactly the same way, we can prove

$$z_n(\alpha) = \alpha + \int_{\alpha}^{1} G_n(\alpha,y)dy$$

$$z_o(\alpha) = 1$$

(4.1) and

$$z_{n+1}(\alpha) = \int_0^{\alpha} t\phi(t)dt + \int_0^{\alpha}(1-t)z_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt$$

$$+ \int_{\alpha}^{1} t z_n\left(\frac{\alpha}{t}\right)\phi(t)dt$$

It follows immediately that for all integers $n \geq o$, the functions $e_n(\alpha)$ satisfy the equations $e_o(\alpha) = 1$.

$$(4.2)\qquad e_{n+1}(\alpha) = \int_0^{\alpha}(1-t)\, e_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt + \int_{\alpha}^{1} t\, e_n\left(\frac{\alpha}{t}\right)dt$$

and the functions $w_n(\alpha)$ satisfy the equations $w_o(\alpha) = 1$.

$$(4.3)\quad w_{n+1}(\alpha) = 2\int_0^{\alpha} t\phi(t)dt + \int_0^{\alpha}(1-t)w_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt$$

$$+ \int_{\alpha}^{1} t w_n\left(\frac{\alpha}{t}\right)\phi(t)dt\,.$$

Theorem 4.16 $h_o(\alpha) = 0$. For all integers $n \geq 0$,

$$h_{n+1}(\alpha) = \int_o^\alpha t^2\phi(t)dt + 2\int_o^\alpha t(1-t)y_n(\tfrac{\alpha-t}{1-t})\phi(t)dt$$
$$+\int_o^\alpha (1-t)^2 h_n(\tfrac{\alpha-t}{1-t})\phi(t)dt + \int_\alpha^1 t^2 h_n(\tfrac{\alpha}{t})\phi(t)\,dt.$$

Proof
$$h_n(\alpha) = \iint_R x^2 dG_n(x,y)$$
$$= \int_o^\alpha x^2 dG_n(x,\alpha)$$
$$= \alpha^2 - 2\int_o^\alpha xG_n(x,\alpha)dx \quad \text{on integration by parts.}$$

So by theorem 4.13,
$$\alpha^2 - h_{n+1}(\alpha) = 2\int_o^\alpha xG_{n+1}(x,\alpha)dx$$
$$= 2\int_o^\alpha \int_o^x xG_n(\tfrac{x-t}{1-t}\ ,\ \tfrac{\alpha-t}{1-t})\phi(t)dt\,dx$$
$$+ 2\int_o^\alpha \int_\alpha^1 xG_n(\tfrac{x}{t}\ ,\ \tfrac{\alpha}{t})\phi(t)dt\,dx.$$

By changing the order of integration, and then changing variables, the right hand side can be written
$$2\int_o^\alpha \int_o^{\frac{\alpha-t}{1-t}} (1-t)[t + u(1-t)]G_n(u, \tfrac{\alpha-t}{1-t})\ \phi(t)\ du\ dt$$
$$+\ 2\int_\alpha^1 \int_o^{\frac{\alpha}{t}} ut^2 G_n(u, \tfrac{\alpha}{t})\ \phi(t)\ du\ dt$$
$$= 2\int_o^\alpha t(1-t)[\tfrac{\alpha-t}{1-t} - y_n(\tfrac{\alpha-t}{1-t})]\phi(t)\ dt \quad +$$

$$+\int_0^{\alpha}(1-t)^2\left[\left(\frac{\alpha-t}{1-t}\right)^2 - h_n\left(\frac{\alpha-t}{1-t}\right)\right]\phi(t)\,dt$$

$$+\int_{\alpha}^{1} t^2\left[\left(\frac{\alpha}{t}\right)^2 - h_n\left(\frac{\alpha}{t}\right)\right]\phi(t)\,dt.$$

The result follows on simplification.

Similarly, one can prove that

$$k_0(\alpha) = 1, \text{ and for all integers } n \geq 0$$

$$k_{n+1}(\alpha) = \int_0^{\alpha} t^2\phi(t)dt + 2\int_0^{\alpha} t(1-t)z_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt \tag{4.4}$$

$$+\int_0^{\alpha}(1-t)^2k_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt + \int_{\alpha}^{1} t^2k_n\left(\frac{\alpha}{t}\right)\phi(t)\,dt.$$

In order to evaluate the product moment $p_n(\alpha)$, and hence $f_n(\alpha)$, we need to know the value of

$$\int_{\alpha}^{1}\int_0^{\alpha} G_n(x,y)\,dx\,dy.$$

Denoting this integral by $q_n(\alpha)$, we have

$$q_{n+1}(\alpha) = \int_{\alpha}^{1}\int_0^{\alpha} G_{n+1}(x,y)dx\,dy$$

$$= \int_{\alpha}^{1}\int_0^{\alpha}\int_0^{x} G_n\left(\frac{x-t}{1-t}, \frac{y-t}{1-t}\right)\phi(t)dtdxdy$$

$$+ \int_{\alpha}^{1}\int_0^{\alpha}\int_y^{1} G_n\left(\frac{x}{t}, \frac{y}{t}\right)\phi(t)\,dtdxdy$$

which, by changing order of integration, and changing variables, can be written

$$\int_0^{\alpha}\int_{\frac{\alpha-t}{1-t}}^{1}\int_0^{\frac{\alpha-t}{1-t}} G_n(u,v)\phi(t)(1-t)^2\,dudvdt$$

$$+\int_{\alpha}^{1}\int_{\frac{\alpha}{t}}^{1}\int_0^{\frac{\alpha}{t}} G_n(u,v)\phi(t)t^2\,dudvdt.$$

So

$$(4.5)\quad q_{n+1}(\alpha) = \int_0^{\alpha}(1-t)^2 q_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt + \int_{\alpha}^{1} t^2 q_n\left(\frac{\alpha}{t}\right)\phi(t)dt\ .$$

Theorem 4.17 $p_0(\alpha) = 0$ and for all integers $k \geq 0$

$$p_{k+1}(\alpha) = \int_0^{\alpha} t^2\phi(t)dt + \int_0^{\alpha} t(1-t)w_k\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt$$

$$+\int_0^{\alpha}(1-t)^2 p_k\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt + \int_{\alpha}^{1} t^2 p_k\left(\frac{\alpha}{t}\right)\phi(t)dt\ .$$

Proof We form the approximating Riemann-Sum for the double integral $\iint_R xydG_k(x,y)$ and rearrange terms.

If $\quad 0 = x_0 < x_1 < x_2 \ldots < x_n = \alpha$

and $\quad \alpha = y_0 < y_1 < y_2 \ldots < y_m = 1$

are partitions of the intervals $[0,\alpha]$ and $[\alpha,1]$ respectively, and ξ_i, η_j, are arbitrary points such that

$$x_{i-1} \leq \xi_i < x_i \quad \text{for } i = 1, 2, \ldots n$$

$$y_{j-1} \leq \eta_j < y_j \quad \text{for } j = 1, 2, \ldots m$$

then the approximating sum

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \xi_i\eta_j[G_k(x_i,y_{j-1}) - G_k(x_i,y_j) - G_k(x_{i-1},y_{i-1} + G_k(x_{i-1},y_j)]$$

can be rearranged and simplified exactly as in the proof of theorem 4.12, to give

$$\sum_{i=1}^{n-1}\sum_{j=1}^{m-1}[\xi_i\,\eta_{j+1} - \xi_i\,\eta_j - \xi_{i+1}\,\eta_{j+1} + \xi_{i+1}\,\eta_j]G_k(x_i\ y_j)$$

$$+ \sum_{j=1}^{m+1}[\xi_n\,\eta_{j+1} - \xi_n\,\eta_j]G_k(\alpha,\ y_j)$$

$$+ \sum_{i=1}^{n-1}[\xi_i\,\eta_1 - \xi_{i+1}\eta_1]G_k(x_i\ \alpha)$$

$$+ \xi_n\eta_1\ G_k(\alpha,\alpha).$$

As the mesh of the partition approaches zero, these terms approach the value

$$-\int_{\alpha}^{1}\int_{o}^{\alpha} G_k(x,y)\ dxdy + \int_{\alpha}^{1}\alpha G_k(\alpha,y)dy - \int_{o}^{\alpha}\alpha G_k(\xi,\alpha)d\xi + \alpha^2\ .$$

So

$$p_k(\alpha) = -q_k(\alpha) + \alpha[z_k(\alpha) - \alpha] - \alpha[\alpha - y\ (\alpha)] + \alpha^2$$

$$p_k(\alpha) = \alpha w_k(\alpha) - q_k(\alpha) - \alpha^2\ .$$

By using the recurrence relations (4,3) and (4,5) for $w_k(\alpha)$ and $q_k(\alpha)$, it is now easy to derive the final result.

We remark that the method used to evaluate $\iint_R xy\ dG_k(x,y)$ may be adapted to finding product moments of higher order.

Finally, by using the three recurrence relations for $h_n(\alpha)$, $k_n(\alpha)$, $p_n(\alpha)$, one can prove easily

<u>Theorem 4.18</u> $f_o(\alpha) = 1$. For all integers $n \geq o$

$$f_{n+1}(\alpha) = \int_{o}^{\alpha}(1-t)^2 f_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt + \int_{\alpha}^{1} t^2 f_n\left(\frac{\alpha}{t}\right)\phi(t)dt.$$

4 - 6 The behaviour of $e_n(\alpha)$

We now attempt to investigate the behaviour of $e_n(\alpha)$ and $v_n(\alpha)$ as $n \to \infty$, and to show how this depends on the density function $\phi(x)$. While it is easy to compute explicitly the values of $\int_0^1 e_n(\alpha)d\alpha$ and $\int_0^1 f_n(\alpha)d\alpha$, giving a global solution , it is much more difficult to consider the sequence $e_n(\alpha)$ for each fixed α. Our main result is that for each α, $0 \leq \alpha \leq 1$

$$e_n(\alpha) \sim \lambda^n \bar{e}(\alpha)$$

where λ is a constant depending on ϕ alone, and $\bar{e}(\alpha)$ is a solution to a homogeneous linear integral equation, which under suitable restrictions, can be solved by an equivalent ordinary differential equation.

We define two numbers λ and μ by the equations

$$\lambda = \int_0^1 (1-t)^2\phi(t)dt + \int_0^1 t^2\phi(t)\,dt$$

$$\mu = \int_0^1 (1-t)^3\phi(t)dt + \int_0^1 t^3\phi(t)dt.$$

The importance of these numbers is apparent by the following result:

Theorem 4.19 For all integers $n \geq 0$

$$\int_0^1 e_n(\alpha)\,d\alpha = \lambda^n$$

$$\int_0^1 f_n(\alpha)\,d\alpha = \mu^n$$

Proof We take the recurrence relation (4.2) for $e_n(\alpha)$, and integrate both sides of the equation with respect to α. This gives

$$\int_0^1 e_{n+1}(\alpha)d\alpha = \int_0^1\int_1^\alpha (1-t)e_n(\tfrac{\alpha-t}{1-t})\,\phi(t)\,dt\,d\alpha$$

$$+ \int_0^1\int_\alpha^1 t\,e_n(\tfrac{\alpha}{t})\,\phi(t)\,dt\,d\alpha$$

$$= \int_0^1\int_t^1 (1-t)\,e_n(\tfrac{\alpha-t}{1-t})\,\phi(t)\,d\alpha\,dt$$

$$+ \int_0^1\int_0^t t\,e_n(\tfrac{\alpha}{t})\,\phi(t)\,d\alpha\,dt$$

$$= \int_0^1\int_0^1 (1-t)^2\,\phi(t)\,e_n(u)\,du\,dt$$

$$+ \int_0^1 t^2\,\phi(t)\,e_n(u)\,du\,dt$$

$$= \lambda \int_0^1 e_n(u)\,du\,.$$

Since $$\int_0^1 e_0(\alpha)\,d\alpha = \int_0^1 d\alpha = 1,$$

$$\int_0^1 e_n(\alpha)\,d\alpha = \lambda^n\,.$$

Similarly, from the recurrence equation for $f_n(\alpha)$ in theorem 4.18,

$$\int_0^1 f_{n+1}(\alpha)\, d\alpha = \int_0^1\int_0^1 (1-t)^3\, \phi(t)\, f_n(u)\, du\, dt$$

$$+\int_0^1\int_0^1 t^3\, \phi(t)\, f_n(u)\, du\, dt$$

$$= \mu \int_0^1 f_n(u)\, du\ ,$$

and since

$$\int_0^1 f_0(u)\, du = 1$$

the proof is complete.

Thus the speed at which $\int_0^1 e_n(\alpha)\, d\alpha$ approaches 0 depends only on λ, that is, only on the first two moments of ϕ. Bounds for λ are provided by the estimates

$$\lambda = \int_0^1 [t^2 + (1-t)^2]\, \phi(t)\, dt \leq \int_0^1 \phi(t)\, dt = 1$$

and

$$\lambda = \int_0^1 [t^2 + (1-t)^2]\, \phi(t)\, dt \geq \int_0^1 \tfrac{1}{2}\, \phi(t)\, dt = \tfrac{1}{2}\ .$$

Moreover, we cannot have equality, for if $\lambda = \frac{1}{2}$,

$\int_0^1 (2t-1)^2\, \phi(t)\, dt = 0$, from which $\phi(t) = 0$ for all t, since ϕ is continuous, and if $\lambda = 1$

$\int_0^1 t(1-t)\, \phi(t)\, dt = 0$, which again gives $\phi(t) = 0$.

But λ can take on any value between $\frac{1}{2}$ and 1 by suitable choice of ϕ. The fact that $\lambda > \frac{1}{2}$ shows that we cannot expect to get a faster rate of convergence to α than that provided by the bisection strategy. The symmetric Beta-distribution

$$\phi(t) = \frac{t^{m-1}(1-t)^{m-1}}{B(m,m)}$$

where m is positive, gives

$$\lambda = \frac{m+1}{2m+1} ,$$

which can be made arbitrarily close to $\frac{1}{2}$ if m is large enough.

If $\phi(t)$ is symmetric about $t = \frac{1}{2}$, then

$$\int_0^1 t\,\phi(t)dt = \frac{1}{2}$$

and $$\lambda = \frac{1}{2} + 2\mu_2 ,$$

where $$\mu_2 = \int_0^1 (t - \tfrac{1}{2})^2\,\phi(t)\,dt$$

is the variance of ϕ. Thus the smaller μ_2 is, the closer λ is to $\frac{1}{2}$, and the more likely it is that abscissa sets are selected near the mid-points of intervals.

The previous theorem suggests that for large n, the sequence $e_n(\alpha)$ should behave like $\lambda^n e(\alpha)$ where $e(\alpha)$ is a fixed function α such that

$$\int_0^1 e(\alpha)\,d\alpha = 1 .$$

It is convenient to normalize the functions $e_n(\alpha)$, by defining a new sequence $\bar{e}_n(\alpha)$ such that

$$\bar{e}_n(\alpha) = \frac{1}{\lambda^n}\, e_n(\alpha).$$

Then, $\bar{e}_0(\alpha) = 1$,

$$(4.6)\qquad \lambda \bar{e}_{n+1}(\alpha) = \int_0^{\alpha} (1-t)\bar{e}_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt + \int_{\alpha}^{1} t\bar{e}_n\left(\frac{\alpha}{t}\right)\phi(t)dt$$

and $$\int_0^1 \bar{e}_n(\alpha)\, d\alpha = 1 , \quad \text{for all integers } n \geq 0.$$

The question of the behaviour of $e_n(\alpha)$ for large n can now be put in another way. Does $\lim_{n \to \infty} \bar{e}_n(\alpha)$ exist? If so, does the limit, $\bar{e}(\alpha)$ say, satisfy the equation

$$\lambda \bar{e}(\alpha) = \int_0^{\alpha} (1-t)\ \bar{e}\ \left(\frac{\alpha-t}{1-t}\right)\ \phi(t)dt + \int_{\alpha}^{1} t\ \bar{e}\ \left(\frac{\alpha}{t}\right)dt \quad ?$$

Regarding the right hand side of this equation as a linear operator, this amounts to asking if the sequence $\bar{e}_n(\alpha)$ converges pointwise to an eigen-function of the linear operator. By choice of λ, this is not a contraction operator in either the L_1-norm or sup-norm sense, so we have to consider other methods of solution.

Henceforth we impose two additional restrictions on $\phi(t)$. These are

$$(4.7)\qquad \text{(i)}\quad \phi(t) \leq M \text{ for all } t,\ 0 \leq t \leq 1 .$$

$$(4.8)\qquad \text{(ii)}\quad \phi(t) \text{ is symmetric about } t = \tfrac{1}{2} , \text{ so that}$$

$$\phi(t) = \phi(1-t) \quad \text{for all } t.$$

We note that $M \geq 1$ since $\int_0^1 \phi(t)dt = 1$, and that

$$(4.9)\qquad \int_0^1 t\ \phi(t)\ dt = \frac{1}{2}$$

$$(4.10)\qquad \lambda = 2\int_0^1 t^2\phi(t)\ dt, \quad \mu = 2\int_0^1 t^3\ \phi(t)\ dt.$$

The second restriction simplifies most of the subsequent calculations. In particular, $\bar{e}_n(\alpha)$ is also symmetric about $\alpha = \frac{1}{2}$ for all n.

The proof that $\lim_{n \to \infty} \bar{e}_n(\alpha)$ exists for each α proceeds by four main stages. First, we show that $\bar{e}_n(\alpha)$ is uniformly bounded for all n and α. Next, that $\int_0^1 t^k \bar{e}_n(t)\, dt$ is a cauchy sequence for each positive integer k. From this it will follow that $\int_0^1 \psi(t)\, \bar{e}_n(t)\, dt$ is a cauchy sequence for all uniformly continuous functions $\psi(t)$. Finally, we can approximate the right hand side of the equation (4.6) by $\int_0^1 \psi(t)\, \bar{e}_n(t) dt$, where ψ depends on α.

Lemma 4.20 $\bar{e}_n(0) = \bar{e}_n(1) = (\frac{1}{2\lambda})^n$ for all positive integers n.

Proof Put $\alpha = 0$ in (4.6).

Then $$\lambda \bar{e}_{n+1}(0) = \int_0^1 t\, \bar{e}_n(0)\, \phi(t)\, dt = \frac{1}{2}\, \bar{e}_n(0) .$$

Theorem 4.21 There exists a real number β, $0 < \beta < \frac{1}{2}$ such that

$$\bar{e}_n(\alpha) \leq \frac{M}{\lambda} \left(\frac{1}{\beta} + \frac{1}{1-\beta} \right)$$

for all α, $0 \leq \alpha \leq 1$, and all non-negative integers n.

Proof We first define β, and then consider separately the cases $0 \leq \alpha \leq \beta$, $\beta \leq \alpha \leq 1-\beta$.

The function $\int_\alpha^1 u\, \phi(u)\, du + \int_0^\alpha (1-u)\, \phi(u)\, du$ is uniformly continuous in α, and has the value $\int_0^1 u\, \phi(u) du = \frac{1}{2}$ if either $\alpha = 0$ or $\alpha = 1$. Thus, since $\lambda > \frac{1}{2}$, there is a β, $0 < \beta < \frac{1}{2}$ such that if either $0 \leq \alpha \leq \beta$ or $1-\beta \leq \alpha \leq 1$, then

$$(4.11)\qquad \int_{\alpha}^{1} u\,\phi(u)\,du \;+\; \int_{0}^{\alpha} (1-u)\phi(u)\,du \;\le\; \lambda\ .$$

Suppose $\beta \le \alpha \le 1-\beta$. Then, by a change of variable in equation (4.6) and using the fact that ϕ is symmetric,

$$\lambda\bar{e}_{n+1}(\alpha) = (1-\alpha)^2 \int_{1-\alpha}^{1} \frac{\bar{e}_n(u)}{u^3}\,\phi\left(\frac{1-\alpha}{u}\right)du + \alpha^2 \int_{\alpha}^{1} \frac{\bar{e}_n(u)}{u^3}\,\phi\left(\frac{\alpha}{u}\right)du$$

$$\le M\left[(1-\alpha)^2 \int_{1-\alpha}^{1} \frac{\bar{e}_n(u)}{u^3}\,du \;+\; \alpha^2 \int_{\alpha}^{1} \frac{\bar{e}_n(u)}{u^3}\,du\right]$$

$$\le M\left[\frac{1}{(1-\alpha)} \int_{1-\alpha}^{1} \bar{e}_n(u)\,du \;+\; \frac{1}{\alpha}\int_{\alpha}^{1} \bar{e}_n(u)\,du\right]$$

$$\le M\left[\frac{1}{1-\alpha} \;+\; \frac{1}{\alpha}\right]\ .$$

So, if $\beta \le \alpha \le 1-\beta$,

$$\bar{e}_{n+1}(\alpha) \;\le\; \frac{M}{\lambda}\left[\frac{1}{\beta} + \frac{1}{1-\beta}\right] \quad \text{for all integers } n \ge 0$$

and hence

$$\bar{e}_n(\alpha) \;\le\; \frac{M}{\lambda}\left[\frac{1}{\beta} + \frac{1}{1-\beta}\right] \quad \text{for all integers } n \ge 0$$

since

$$\frac{M}{\lambda}\left[\frac{1}{\beta} + \frac{1}{1-\beta}\right] \ge 4 \ge \bar{e}_0(\alpha)\ .$$

We prove that $\bar{e}_n(\alpha)$ satisfies the same inequality for all α by induction on n. This is trivial if n = 0, since $\bar{e}_0(\alpha)=1$. Suppose for some positive integer k,

$$\bar{e}_k(\alpha) \;\le\; \frac{M}{\lambda}\left(\frac{1}{\beta} + \frac{1}{1-\beta}\right) \quad \text{for all } \alpha\ .$$

Then, if $0 \le \alpha \le \beta$,

$$\lambda \bar{e}_{k+1}(\alpha) = \int_0^{\alpha} (1-t)\, \bar{e}_k\left(\frac{\alpha-t}{1-t}\right) \phi(t)dt + \int_{\alpha}^{1} t\, \bar{e}_k\left(\frac{\alpha}{t}\right) \phi(t)\, dt$$

$$\leq \frac{M}{\lambda}\left[\frac{1}{\beta} + \frac{1}{1-\beta}\right]\left[\int_0^{\alpha} (1-t)\phi(t)\, dt + \int_{\alpha}^{1} t\phi(t)dt\right]$$

$$\leq \frac{M}{\lambda}\left[\frac{1}{\beta} + \frac{1}{1-\beta}\right]\lambda \quad \text{by equation (4.11).}$$

So

$$\bar{e}_{k+1}(\alpha) \leq \frac{M}{\lambda}\left[\frac{1}{\beta} + \frac{1}{1-\beta}\right] \quad \text{for} \quad 0 \leq \alpha \leq \beta$$

and, since $\bar{e}_{k+1}$ is symmetric,

$$\bar{e}_{k+1}(\alpha) \leq \frac{M}{\lambda}\left[\frac{1}{\beta} + \frac{1}{1-\beta}\right] \text{ for all } \alpha,\ 0 \leq \alpha \leq 1 .$$

We prove now that the sequence $\int_0^1 t^k\, \bar{e}_n(t)dt$ is a cauchy sequence for each non-negative integer k. To do this, it is desirable to introduce some new notation.

Definition 4.22 Let p be an arbitrary positive integer, and put $\delta_n(\alpha) = \bar{e}_{n+p}(\alpha) - \bar{e}_n(\alpha)$ for each non-negative integer n. Let k be any non-negative integer, and put

$$B_{n,k} = \int_0^1 t^k \delta_n(t)\, dt.$$

Theorem 4.23 Choose E, E fixed, $E \geq \sup_{n,\alpha} |\delta_n(\alpha)|$. This is possible by theorem 4.21.

Since

$$\frac{\mu}{\lambda} = \frac{2\int_0^1 u^3\, \phi(u)du}{2\int_0^1 u^2\, \phi(u)\, du} < 1$$

there is a number γ such that $\frac{\mu}{\lambda} < \gamma < 1$.

Define a sequence A_k for each integer $k > 0$ by the equations

$$A_o = 1$$

$$A_k = 2E + \frac{\lambda}{\lambda\gamma-\mu} \sum_{r=0}^{k-1} \binom{k}{r} A_r \tag{4.12}$$

$$\text{where } \binom{k}{r} = \frac{k!}{r!(k-r)!}$$

Then $|B_{n,k}| \leq A_k \gamma^n$ for all k, n. In particular;

$$\lim_{n \to \infty} |B_{n,k}| = 0 .$$

Proof We use induction on k.

First, if k = 0,

$$B_{n,o} = \int_o^1 \delta_n(t)\, dt$$

$$= \int_o^1 \bar{e}_{n+p}(t)\, dt - \int_o^1 \bar{e}_n(t)\, dt$$

$$= 1 - 1$$

$$= 0 .$$

So the result is trivial for k = 0.

Assume that the theorem is proved for all integers r, $0 \leq r \leq k-1$, where $k \geq 1$. From equation 4.6 and definition 4.22, it follows by linearity that

$$\lambda\delta_{n+1}(\alpha) = \int_o^\alpha (1-t)\, \delta_n\left(\frac{\alpha-t}{1-t}\right) \phi(t)\, dt + \int_\alpha^1 t\, \delta_n\left(\frac{\alpha}{t}\right)\phi(t)dt$$

for all integers $n \geq 0$.

Multiplying each side of this equation by α^k and integrating, gives

$$\lambda B_{n+1,k} = \int_o^1 \int_o^\alpha \alpha^k (1-t)\delta_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dtd\alpha + \int_o^1 \int_\alpha^1 \alpha^k t\delta_n\left(\frac{\alpha}{t}\right)\phi(t)\, dtd\alpha$$

which in turn, by putting

$$1-u = \frac{\alpha-t}{1-t}, \quad 1-v = t \quad \text{in the first integral,}$$

$$u = \frac{\alpha}{t}, \quad v = t \quad \text{in the second integral,}$$

gives

$$\lambda B_{n+1,k}$$

$$= \int_0^1 \int_0^1 (1-uv)^k u^2 \, \delta_n(v) \, \phi(u) \, du \, dv + \int_0^1 \int_0^1 v^k u^{k+2} \delta_n(v) \, \phi(u) \, du dv$$

$$= \sum_{r=0}^{k} \binom{k}{r} (-1)^r \int_0^1 \int_0^1 u^{r+2} v^r \delta_n(v) \, \phi(u) du dv + \int_0^1 \int_0^1 v^k u^{k+2} \delta_n(v) \phi(u) du dv$$

$$= \sum_{r=0}^{k} \binom{k}{r} (-1)^r B_{n,r} \int_0^1 u^{r+2} \phi(u) du + B_{n,k} \int_0^1 u^{k+2} \phi(u) du$$

$$= [1 + (-1)^k] \, B_{n,k} \int_0^1 u^{k+2} \, \phi(u) \, du + \sum_{r=0}^{k-1} \binom{k}{r} (-1)^r B_{n,r} \int_0^1 u^{r+2} \phi(u) du$$

Therefore

$$\lambda |B_{n+1,k}|$$

$$\leq 2|B_{n,k}| \int_0^1 u^{k+2} \, \phi(u) du + \sum_{r=0}^{k-1} \binom{k}{r} |B_{n,r}| \int_0^1 u^{r+2} \, \phi(u) du$$

$$\leq 2|B_{n,k}| \int_0^1 u^3 \, \phi(u) du + \sum_{r=0}^{k-1} \binom{k}{r} |B_{n,r}| \int_0^1 u^2 \, \phi(u) du$$

$$= \mu |B_{n,k}| + \frac{1}{2} \lambda \sum_{r=0}^{k-1} \binom{k}{r} |B_{n,r}|.$$

Hence

$$|B_{n+1,k}| \leq \frac{\mu}{\lambda} |B_{n,k}| + \frac{1}{2} \sum_{r=0}^{k-1} \binom{k}{r} |B_{n,r}| .$$

By the induction hypothesis and equation (4.12), this implies that

$$|B_{n+1,k}| \leq \frac{\mu}{\lambda} |B_{n,k}| + \frac{1}{2} \gamma^n \sum_{r=0}^{k-1} \binom{k}{r} A_r$$

i.e.

$$|B_{n+1,k}| \leq \frac{\mu}{\lambda} |B_{n,k}| + \frac{1}{2} \gamma^n (A_k - 2E)(\gamma - \frac{\mu}{\lambda}) .$$

By considering this inequality for all integer values of n from 0 upwards, it follows that

$$|B_{n+1,k}| \leq (\frac{\mu}{\lambda})^{n+1} |B_{o,k}| + \frac{1}{2}(\gamma - \frac{\mu}{\lambda})(A_k - 2E) \sum_{m=0}^{n} \gamma^m (\frac{\mu}{\lambda})^{n-m}$$

$$\leq \gamma^{n+1} |B_{o,k}| + \frac{1}{2} (A_k - 2E)[\gamma^{n+1} - (\frac{\mu}{\lambda})^{n+1}]$$

$$\leq \gamma^{n+1} E + \frac{1}{2} (A_k - 2E) \gamma^{n+1}$$

$$\leq A_k \gamma^{n+1} ,$$

where the inequality $|B_{o,k}| \leq E$ follows from the fact that

$$|B_{o,k}| = |\int_o^1 t^k \delta_o(t) dt| \leq E \int_o^1 t^k dt \leq E.$$

Hence $|B_{n,k}| \leq A_k \gamma^n$ for all k and all n, and so the induction is complete.

Corollary 4.24

$$\lim_{n \to \infty} |\int_o^1 p(t)\ \delta_n(t)\ dt| = 0,$$

where p(t) is any polynomial in t.

Proof p(t) is a sum of the form $\sum_{r=0}^{k} a_r t^r$ where the a_r are constants.

Theorem 4.25 If $\psi(t)$ is any real uniformly continuous function defined for $0 \leq t \leq 1$, then

$$\lim_{n \to \infty} \left| \int_0^1 \psi(t)\, \delta_n(t)\, dt \right| = 0$$

Proof For any $\epsilon > 0$, there exists, by the Weierstrass approximation theorem, a polynomial $p_\epsilon(t)$ such that

$$|p_\epsilon(t) - \psi(t)| < \epsilon \quad \text{for} \quad 0 \leq t \leq 1.$$

Hence

$$\left| \int_0^1 \psi(t)\, \delta_n(t) dt \right| = \left| \int_0^1 [\psi(t) - p_\epsilon(t) + p_\epsilon(t)] \delta_n(t) dt \right|$$

$$\leq \left| \int_0^1 [\psi(t) - p_\epsilon(t)] \delta_n(t) dt \right| + \left| \int_0^1 p_\epsilon(t) \delta_n(t) dt \right|$$

$$\leq \epsilon \int_0^1 |\delta_n(t)| dt + \left| \int_0^1 p_\epsilon(t) \delta_n(t) dt \right|$$

$$\leq \epsilon E + \left| \int_0^1 p_\epsilon(t) \delta_n(t)\, dt \right|$$

Since ϵ is arbitrary, the result follows on letting $n \to \infty$, by corollary 4.24.

Theorem 4.26 The sequence $\delta_n(\alpha)$ converges uniformly to zero on the interval $0 \leq \alpha \leq 1$.

Proof As in theorem 4.21, we must consider first the case $0 < \beta \leq \alpha \leq 1-\beta < 1$, for some suitable β, and then extend the result to cover the case $0 \leq \alpha \leq \beta$. The first part of the proof rests on the equation

$$\lambda\delta_{n+1}(\alpha) = \alpha^2 \int_{\alpha}^{1} \frac{\delta_n(t)}{t^3}\,\phi(\tfrac{\alpha}{t})dt + (1-\alpha)^2 \int_{1-\alpha}^{1} \frac{\delta_n(t)}{t^3}\,\phi(\tfrac{1-\alpha}{t})\,dt,$$

in which $\frac{1}{t^3}\,\phi(\frac{\alpha}{t})$ and $\frac{1}{t^3}\,\phi(\frac{1-\alpha}{t})$ are approximated by suitable continuous functions. Theorem 4.25 can then be applied.

Let $\bar{\epsilon}$ be given, $\bar{\epsilon} > 0$. Choose ϵ, $\epsilon > 0$, such that

(i) $\epsilon < \lambda\bar{\epsilon}\,(\gamma\lambda - \frac{1}{2}) \,/\, 16\,E$

(ii) $\epsilon < \beta$

(iii) $\frac{M\epsilon E}{\lambda}\,(\,\frac{1}{\beta} + \frac{1}{1-\beta}\,) < \frac{\bar{\epsilon}}{2}\,(\gamma\lambda - \frac{1}{2}\,)$

where γ and β are chosen such that $\frac{1}{2\lambda} < \gamma < 1$, $0 < \beta < \frac{1}{2}$, and

(4.13) $$0 < \int_{0}^{\beta} (1-t)\phi(t)dt + \int_{0}^{2\beta} t\phi(t)dt \le \frac{1}{2}\,(\gamma\lambda - \frac{1}{2}\,).$$

These inequalities are possible, since $\lambda > \frac{1}{2}$, and the integral is increasing in β.

For each α, $\beta \le \alpha \le 1-\beta$, define the two functions $\psi_1(t,\alpha)$ and $\psi_2(t,\alpha)$, by the equations

$$\psi_1(t,\alpha) = \begin{cases} 0, \text{ if } 0 \le t \le \alpha-\epsilon \\ \dfrac{[t-(\alpha-\epsilon)]\phi(1)}{\epsilon\alpha^3}, \text{ if } \alpha-\epsilon \le t \le \alpha \\ \dfrac{1}{t^3}\,\phi(\tfrac{\alpha}{t}), \text{ if } \alpha \le t \le 1. \end{cases}$$

$$\psi_2(t,\alpha) = \begin{cases} 0, \text{ if } 0 \le t \le 1-\alpha-\epsilon \\ \dfrac{[t-(1-\alpha-\epsilon)]\phi(1)}{\epsilon(1-\alpha)^3}, \text{ if } 1-\alpha-\epsilon \le t \le 1-\alpha \\ \dfrac{1}{t^3}\,\phi(\tfrac{1-\alpha}{t}), \text{ if } 1-\alpha \le t \le 1. \end{cases}$$

Both $\psi_1(t,\alpha)$ and $\psi_2(t,\alpha)$ are uniformly continuous on the rectangle $\beta \leq \alpha \leq 1-\beta$, $0 \leq t \leq 1$. So there is $\delta > 0$, such that if both α and α_i belong to the closed interval $[\beta, 1-\beta]$, and $|\alpha-\alpha_i| < \delta$, then

$$|\psi_1(t,\alpha) - \psi_1(t,\alpha_i)| < \epsilon$$

and

$$|\psi_2(t,\alpha) - \psi_2(t,\alpha_i)| < \epsilon$$

for all t, $0 \leq t \leq 1$.

Choose a finite set of points α_i, $i = 1, 2, \ldots q$, such that if $\beta \leq \alpha \leq 1-\beta$, there is an α_i such that $|\alpha - \alpha_i| < \delta$.

If $\beta \leq \alpha \leq 1-\beta$, then

$$\lambda\delta_{n+1}(\alpha) = \alpha^2 \int_{\alpha}^{1} \psi_1(t,\alpha)\delta_n(t)dt + (1-\alpha)^2 \int_{1-\alpha}^{1} \psi_2(t,\alpha)\delta_n(t)dt$$

$$= \alpha^2 \int_{0}^{1} \psi_1(t,\alpha)\delta_n(t)dt + (1-\alpha)^2 \int_{0}^{1} \psi_2(t,\alpha)\delta_n(t)\,dt$$

$$- \alpha^2 \int_{\alpha-\epsilon}^{\alpha} \frac{[t-(\alpha-\epsilon)]}{\epsilon\alpha^3}\,\phi(1)\,\delta_n(t)\,dt$$

$$- (1-\alpha)^2 \int_{1-\alpha-\epsilon}^{1-\alpha} \frac{[t-(1-\alpha-\epsilon)]}{\epsilon(1-\alpha)^3}\,\phi(1)\delta_n(t)\,dt.$$

So

$$\lambda|\delta_{n+1}(\alpha)| \leq \left|\int_{0}^{1} \psi_1(t,\alpha)\delta_n(t)dt\right| + \left|\int_{0}^{1} \psi_2(t,\alpha)\delta_n(t)\,dt\right|$$

$$+ \frac{EM}{\epsilon\alpha} \int_{\alpha-\epsilon}^{\alpha} [t-(\alpha-\epsilon)]\,dt$$

$$+ \frac{EM}{\epsilon(1-\alpha)} \int_{1-\alpha-\epsilon}^{1-\alpha} [t-(1-\alpha-\epsilon)]\,dt \leq$$

$$\leq \sum_{j=1}^{2} \left| \int_0^1 [\psi_j(t,\alpha) - \psi_j(t,\alpha_i) + \psi_j(t,\alpha_i)]\, \delta_n(t)\, dt \right|$$

$$+ \frac{E\,M\,\epsilon}{2} \left(\frac{1}{\alpha} + \frac{1}{1-\alpha} \right) \quad \text{where } |\alpha-\alpha_i| < \delta$$

$$\leq 2\epsilon \int_0^1 |\delta_n(t)|dt + \left| \int_0^1 \psi_1(t,\alpha_i)\delta_n(t)\, dt \right|$$

$$+ \left| \int_0^1 \psi_2(t,\alpha_i)\, \delta_n(t)\, dt \right| + \frac{E\,M\,\epsilon}{2} \left(\frac{1}{\beta} + \frac{1}{1-\beta} \right)$$

$$\leq 2\,\epsilon\,E + \frac{\lambda\bar{\epsilon}}{4}(\gamma\lambda - \tfrac{1}{2})$$

$$+ \left| \int_0^1 \psi_1(t,\alpha_i)\delta_n(t)\, dt \right| + \left| \int_0^1 \psi_2(t,\alpha_i)\delta_n(t)dt \right|.$$

By theorem 4.25, and the fact that there are only a finite number of α_i, there exists an integer m_o, such that if $n \geq m_o$, then

$$\left| \int_0^1 \psi_1(t,\alpha_i)\delta_n(t)dt \right| + \left| \int_0^1 \psi_2(t,\alpha_i)\delta_n(t)dt \right|$$

$$< \frac{\lambda\,\bar{\epsilon}}{8} \left(\gamma\lambda - \frac{1}{2} \right) \quad \text{for each } i,\ i = 1,\ 2,\ 3,\ldots\ q.$$

But then,

$$|\delta_{n+1}(\alpha)| \leq \frac{1}{2}\ \bar{\epsilon}\ (\gamma\lambda - \tfrac{1}{2}), \quad \text{for all } n \geq m_o .$$

We wish to show that there is a positive integer p, such that $n \geq m_o + p$, then $|\delta_n(\alpha)| \leq \epsilon$ for all α, $0 \leq \alpha \leq 1$. It remains to consider the case $0 \leq \alpha \leq \beta$.

Define $B_n = \sup_{\alpha} |\delta_n(\alpha)|$, $0 < \alpha \leq \beta$, for all positive integers n. We claim that if $B_n > \frac{\bar{\epsilon}}{2}$, $n \geq m_o + 1$, then the sequence B_n decreases until $B_{n+p} \leq \frac{\bar{\epsilon}}{2}$ for some p.

Assume that $0 < \alpha \le \beta$, $n \ge m_o+1$, and $B_n > \frac{\bar{\epsilon}}{2}$. Then since

$$\lambda\, \delta_{n+1}(\alpha) = \alpha^2 \int_{\alpha}^{1} \frac{\delta_n(t)}{t^3}\, \phi(\tfrac{\alpha}{t})\, dt + (1-\alpha)^2 \int_{1-\alpha}^{1} \frac{\delta_n(t)}{t^3}\, \phi(\tfrac{1-\alpha}{t})dt,$$

$$\lambda|\delta_{n+1}(\alpha)| \le \alpha^2 \int_{1-\beta}^{1} \frac{B_n}{t^3}\, \phi(\tfrac{\alpha}{t})dt + \alpha^2 \int_{\beta}^{1-\beta} \frac{\bar{\epsilon}(\gamma\lambda - \frac{1}{2})}{t^3}\, \phi(\tfrac{\alpha}{t})dt$$

$$+ \alpha^2 \int_{\alpha}^{\beta} \frac{B_n}{t^3}\, \phi(\tfrac{\alpha}{t})\, dt + (1-\alpha)^2 \int_{1-\alpha}^{1} \frac{B_n}{t^3}\, \phi(\tfrac{1-\alpha}{t})\, dt$$

$$\le B_n \int_{\alpha}^{\frac{\alpha}{1-\beta}} u\, \phi(u)dt + \tfrac{1}{2}\, \bar{\epsilon}(\gamma\lambda - \tfrac{1}{2}) \int_{\frac{\alpha}{1-\beta}}^{\frac{\alpha}{\beta}} u\phi(u)du$$

$$+ B_n \int_{\frac{\alpha}{\beta}}^{1} u\phi(u)du + B_n \int_{1-\alpha}^{1} u\phi(u)du$$

$$\le B_n \int_{o}^{2\alpha} u\phi(u)du + \tfrac{1}{2}\, \bar{\epsilon}\, (\gamma\lambda - \tfrac{1}{2}) \int_{o}^{1} u\phi(u)du$$

$$+ B_n \int_{o}^{1} u\phi(u)du + B_n \int_{o}^{\alpha} (1-u)\, \phi\, (u)du$$

$$\le B_n[\int_{o}^{2\beta} u\phi(u)du + \tfrac{1}{2}\, (\gamma\lambda - \tfrac{1}{2}) + \tfrac{1}{2} + \int_{o}^{\beta} (1-u)\phi(u)du]$$

So by assumption and choice of β,

$$\lambda|\delta_{n+1}(\alpha)| \le B_n[(\gamma\lambda - \tfrac{1}{2}) + \tfrac{1}{2}]$$

$$\le \gamma\, \lambda\, B_n\, .$$

It follows that $B_{n+1} \leq \gamma B_n$, and hence if $B_{m_o+1} > \frac{1}{2}\bar{\epsilon}$, the sequence B_{m_o+2}, B_{m_o+3}, ... decreases until $B_{m_o+p} \leq \frac{1}{2}\bar{\epsilon}$, for some p.

Finally, the equation

$$\lambda\, \delta_{n+1}(\alpha) = \int_0^{\alpha} (1-t)\, \delta_n\left(\frac{\alpha-t}{1-t}\right) \phi(t)\, dt + \int_{1-\delta}^{1} t\, \delta_n\left(\frac{\alpha}{t}\right) \phi(t)\, dt$$

implies that

$$\lambda \sup_{0 \leq \alpha \leq 1} |\delta_{n+1}(\alpha)| \leq \sup_{0 \leq \alpha \leq 1} |\delta_n(\alpha)| ,$$

and so $B_{n+1} \leq \frac{1}{\lambda} \sup_{0 \leq \alpha \leq 1} |\delta_n(\alpha)| \leq 2 \sup_{0 \leq \alpha \leq 1} |\delta_n(\alpha)|$

for all n. It follows that $B_n \leq \bar{\epsilon}$ for all $n \geq m_o + p$.

Hence $|\delta_n(\alpha)| \leq \bar{\epsilon}$ for all $n \geq m_o + p$, and for all α, $0 < \alpha < 1$.

By lemma 4.18, this holds also for $\alpha = 0$ and $\alpha = 1$. This completes the proof.

Corollary 4.27. Since $\delta_n(\alpha) = \bar{e}_{n+p}(\alpha) - \bar{e}_n(\alpha)$, where p is an arbitrary positive integer, the sequence $\bar{e}_n(\alpha)$ for each α is a cauchy sequence. So

$\bar{e}(\alpha) = \lim_{n \to \infty} \bar{e}_n(\alpha)$ exists, and moreover, the convergence is uniform.

Lemma 4.28 If the sequence D_n is defined by $D_o = 0$,

$$D_{n+1} = \frac{1}{\lambda} (D_n + C\, M)$$

for each integer $n \geq 0$, where $C = \sup_{n,\alpha} \bar{e}_n(\alpha)$, then

$$|\bar{e}_n(\alpha) - \bar{e}_n(\beta)| \leq D_n|\alpha - \beta|$$

for any α,β in the interval $[0,1]$.

Proof We use induction on n.

The lemma is true for $n = 0$, since $\bar{e}_n(\alpha) = \bar{e}_n(\beta) = 1$.

If $$|\bar{e}_n(\alpha) - \bar{e}_n(\beta)| \leq D_n|\alpha-\beta|$$

for some n, and $\alpha > \beta$, then

$$\lambda|\bar{e}_{n+1}(\alpha) - \bar{e}_{n+1}(\beta)| = \Big|\int_0^\alpha (1-t)\,\bar{e}_n(\tfrac{\alpha-t}{1-t})\,\phi(t)dt$$

$$-\int_0^\beta (1-t)\bar{e}_n(\tfrac{\beta-t}{1-t})\phi(t)dt + \int_\alpha^1 t\bar{e}_n(\tfrac{\alpha}{t})\phi(t)dt - \int_\beta^1 t\bar{e}_n(\tfrac{\beta}{t})\phi(t)dt\Big|$$

$$\leq \Big|\int_\beta^\alpha (1-t)\bar{e}_n\,(\tfrac{\alpha-t}{1-t})\phi(t)dt\Big| + \int_0^\beta (1-t)\phi(t)\,|\bar{e}_n(\tfrac{\beta-t}{1-t}) - \bar{e}_n(\tfrac{\alpha-t}{1-t})|\,dt$$

$$+ \Big|\int_\beta^\alpha t\bar{e}_n(\tfrac{\beta}{t})\phi(t)dt\Big| + \int_\alpha^1 t\phi(t)\,|\,\bar{e}_n(\tfrac{\alpha}{t}) - \bar{e}_n(\tfrac{\beta}{t})|\,dt$$

$$\leq C\int_\beta^\alpha (1-t)\phi(t)dt + D_n|\alpha-\beta|\int_0^\beta \phi(t)\,dt$$

$$+ C\int_\beta^\alpha t\,\phi(t)dt + D_n|\alpha-\beta|\int_\alpha^1 \phi(t)\,dt$$

$$\leq C\,M|\alpha-\beta| + D_n|\alpha-\beta|$$

$$= \lambda\, D_{n+1}|\alpha-\beta|\,.$$

Corollary 4.29 $\bar{e}_n(\alpha)$ is uniformly continuous on $0 \leq \alpha \leq 1$ for each n.

Corollary 4.30 $\bar{e}(\alpha) = \lim_{n \to \infty} \bar{e}_n(\alpha)$ is uniformly continuous.

Proof $\bar{e}(\alpha)$ is the limit of a uniformly convergent sequence of continuous functions.

Corollary 4.31 $\bar{e}(\alpha)$ satisfies the integral equation

$$\lambda \bar{e}(\alpha) = \int_0^\alpha (1-t)\, \bar{e}\left(\frac{\alpha-t}{1-t}\right) \phi(t)dt + \int_\alpha^1 t\, \bar{e}\left(\frac{\alpha}{t}\right) \phi(t)dt.$$

Proof

$$\lambda \bar{e}(\alpha) = \lambda \lim_{n \to \infty} \bar{e}_{n+1}(\alpha)$$

$$= \lim_{n \to \infty} \int_0^\alpha (1-t)\bar{e}_n\left(\frac{\alpha-t}{1-t}\right)\phi(t)dt + \lim_{n \to \infty} \int_\alpha^1 t\bar{e}_n\left(\frac{\alpha}{t}\right)\phi(t)dt,$$

and a change of order of limit and integration is permissible by uniform convergence.

The next lemma shows that to compute $\bar{e}(\alpha)$, one may differentiate the integral equation and solve the resulting ordinary differential equation, so long as ϕ is differentiable a sufficient number of times. The proof is omitted, since it follows from the classical methods for differentiation under the integral sign.

Lemma 4.32 If $\phi(t)$ is N times continuously differentiable on $[0,1]$, then $\bar{e}(\alpha)$ is N times differentiable in the open interval $(0,1)$.

4 - 7 The rectangular density function

As an example, let us consider briefly the case $\phi(t) \equiv 1$. It is easily checked that

$$\lambda = 2 \int_0^1 t^2 \, \phi(t) \, dt = \frac{2}{3} ,$$

and consequently

$$\frac{2}{3} \bar{e}(\alpha) = \alpha^2 \int_\alpha^1 \frac{\bar{e}(t)}{t^3} \, dt + (1-\alpha)^2 \int_{1-\alpha}^1 \frac{\bar{e}(t)}{t^3} \, dt,$$

for $0 < \alpha < 1$.

By differentiating twice, we have

$$2\alpha(1-\alpha)\bar{e}''(\alpha) + (1-2\alpha)\bar{e}'(\alpha) + 2\bar{e}(\alpha) = 0,$$

and if we put

$$\bar{e}(\alpha) = \sqrt{\alpha(1-\alpha)} \; g(\alpha) \text{ , then}$$

$$2\alpha(1-\alpha)g''(\alpha) + 3(1-2\alpha)g'(\alpha) = 0.$$

The differential equation for g has two linearly independent solutions only, namely

$$g(\alpha) = \text{constant, and}$$

$$g(\alpha) = k \operatorname{Cot} \phi$$

where k is a constant, and $\alpha = \operatorname{Sin}^2 \frac{\phi}{2}$.

The second of these solutions yields a solution $\bar{e}(\alpha)$ of the differential equation which becomes unbounded in a neighbourhood of $\alpha = 0$, or $\alpha = 1$. This possibility must be rejected.

So $g(\alpha) = c$ say,

and $\bar{e}(\alpha) = c\sqrt{\alpha(1-\alpha)}$.

To satisfy the requirement $\int_0^1 \bar{e}(\alpha) d\alpha = 1$, we must take

$c = \frac{8}{\pi}$, and so finally,

$$\lim_{n \to \infty} \bar{e}_n(\alpha) = \frac{8}{\pi} \sqrt{\alpha(1-\alpha)} \qquad \text{for all } \alpha,\ 0 \leq \alpha \leq 1.$$

Hence $$e_n(\alpha) \sim \frac{8}{\pi} \left(\frac{2}{3}\right)^n \sqrt{\alpha(1-\alpha)} \ .$$

This result was suggested by E.N. Gilbert of Bell Telephone Laboratories.

It is of interest to compute the variance $v_n(\alpha)$ for this case. Putting $\phi(t) = 1$, and using theorem 4.18, one shows first by induction on n that

$$f_n(\alpha) = a_n + b_n\, \alpha(1-\alpha)$$

where

$$a_0 = 1 \qquad a_{n+1} = \tfrac{1}{3} a_n \qquad n = 0,\ 1,\ 2,\ \ldots$$
$$b_0 = 0 \qquad b_{n+1} = \tfrac{1}{2} b_n + a_n \qquad n = 0,\ 1,\ 2,\ \ldots\ .$$

Hence $$f_n(\alpha) = \left(\tfrac{1}{3}\right)^n + 6\left[\left(\tfrac{1}{2}\right)^n - \left(\tfrac{1}{3}\right)^n\right] \alpha\,(1-\alpha).$$

It follows that

$$\begin{aligned} v_n(\alpha) &= f_n(\alpha) - [e_n(\alpha)]^2 \\ &= \left(\tfrac{1}{3}\right)^n + 6\left[\left(\tfrac{1}{2}\right)^n - \left(\tfrac{1}{3}\right)^n\right] \alpha(1-\alpha) - [e_n(\alpha)]^2 \\ &\sim \left(\tfrac{1}{3}\right)^n + 6\left[\left(\tfrac{1}{2}\right)^n - \left(\tfrac{1}{3}\right)^n\right] \alpha(1-\alpha) - \frac{64}{\pi^2} \left(\tfrac{4}{9}\right)^n \alpha(1-\alpha) \end{aligned}$$

So $$v_n(\alpha) \sim \frac{6\alpha(1-\alpha)}{2^n} \ .$$

Similarly, one can show by induction on n that if $\phi = 1$,

$$w_n(\alpha) = 2\alpha + \frac{1-2\alpha}{2^n} \qquad n = 0,\ 1,\ 2,\ \ldots$$

which suggests that the mean value of l_n and r_n could be used to estimate α if n is sufficiently large.

However, the variance of $w_n(\alpha)$ is too large for an effective estimate, since the variance of $w_n(\alpha)$ is $V_n(\alpha)$, where

$$V_n(\alpha) = E(l_n + r_n)^2 - [E(l_n + r_n)]^2$$

$$= f_n(\alpha) + 4p_n(\alpha) - [w_n(\alpha))]^2 .$$

One can show easily that if $\phi = 1$, then

$$p_n(\alpha) = \alpha^2[1 - \tfrac{1}{2}n] .$$

Combining this result with the known values for $f_n(\alpha)$ and $w_n(\alpha)$ gives the variance of $w_n(\alpha)$,

$$V_n(\alpha) = \frac{2\alpha(1-\alpha)}{2^n} + \frac{1 - 6\alpha(1-\alpha)}{3^n} - \frac{(1 - 2\alpha)^2}{4^n}$$

i.e.

$$V_n(\alpha) \sim \frac{2\alpha(1-\alpha)}{2^n}$$

Hence $\sqrt{V_n(\alpha)} > e_n(\alpha)$ if n is large enough.

CHAPTER 5

LOCALIZATION OF ZEROES OF DERIVATIVES

5 - 1 General Results

We consider the problem of localizing the unique zero of the kth order derivative of a suitable function defined on an interval.

It is convenient to change some of the notation used in previous chapter. Let D be the unit interval $[0,1]$, and let k be any positive integer. Let F_k be the class of all real-valued functions f defined on D which possess continuous monotonic kth order derivatives $f^{(k)}(x)$, such that $f^{(k)}(0) < 0 < f^{(k)}(1)$. The functional is defined by the unique zero of $f^{(k)}$.

In place of the exponents ρ, ρ_N and $\rho_N(M)$, we use $\rho(k)$, $\rho_N(k)$, and $\rho_N(k;M)$ respectively. See section 3-3.

Lemma 5.1 For each k, the functional is localizable with respect to a (k+1)th order test function. Moreover, $\rho(k) \geq 2^{\frac{1}{k+1}}$

Proof Let $H = [a,b]$ be any closed sub-interval of D. Select k+1 points $x_1, x_2, \ldots x_{k+1}$ in H such that

$$a < x_1 < x_2 < x_3 \ldots < x_k < x_{k+1} < b.$$

Define the test function T on the set $D^{k+1} \times E^{k+1}$ by:

$$T = T(x_1, x_2, \ldots, x_{k+1}, f(x_1), f(x_2), \ldots, f(x_{k+1}))$$

is the sign of the kth difference quotient of f at the points

$x_1, x_2, \ldots x_{k+1}$ where $f \in F_k$. The kth difference quotient may be defined as the coefficient of t^k in the unique polynomial $\phi_k(t)$ of degree $\leq k$ which assumes the value $f(x_i)$ at $t = x_i$, $i = 1, 2, 3, \ldots (k+1)$. See for example van der Waerden [9], on page 67.

It follows immediately from the definition of F_k that if $f \in F_k$ and $S(f) \in H$, then $S(f) < x_1$ implies $T > 0$, and $S(f) > x_{k+1}$ implies $T < 0$. Hence

$$S(f) \leq x_{k+1} \quad \text{if} \quad T \geq 0$$

$$\text{and} \qquad S(f) \geq x_1 \quad \text{if} \quad T \leq 0 \ .$$

Thus, no matter what value T assumes, S(f) can always be restricted to a subinterval of H of length at most $\text{Max}\{ x_{k+1} - a, b - x_1 \}$, and if, for some $\epsilon > 0$, the distance from both x_1 and x_{k+1} to the midpoint of H is $\frac{\epsilon}{2}$, the subinterval of H has length at most $(\frac{1}{2} + \epsilon)(b - a)$.

It follows from lemma 3.2 that the functional is localizable. Moreover, if the length of $D = d_o = 1$, then there is a strategy M, such that

$$d_{n(k+1)}(M) = (\tfrac{1}{2} + \epsilon)^n .$$

But this implies that

$$\rho_{n(k+1)}(k,M) = [(\tfrac{1}{2} + \epsilon)^{-n}]^{\frac{1}{n(k+1)}} ,$$

from which $$\rho(k) \geq 2^{\frac{1}{k+1}} .$$

We can include the case k = 0 in the above, by agreeing that the derivative of order zero of a function at a point is the value of the function at that point. This gives the bisection problem, for which $\rho(0) = 2$.

The case k = 1 is done already, since F_1 is a subclass of the unimodal functions considered in section 1-3. The above inequality $\rho(1) \geq \sqrt{2} = 1.414...$ is a poor one, since

$$\rho(1) = (\sqrt{5} + 1)/2 = 1.618... .$$

The cases k = 0, k = 1, and k = 2 are all discussed in detail by Kiefer [6], who gives a complete analysis of the optimal strategies for these three problems. In particular, he shows that $\rho(2) = \sqrt{2}$.

We wish to obtain estimates for $\rho(k)$, where k is any positive integer. The methods employed are very similar to those used in the proof of theorem 1.1.

Theorem 5.2 Let N be a positive integer. Define $L_N(k)$ to be the largest real number such that if the domain of F_k is the closed interval $D = [0, L_N(k)]$, then there is a strategy M for which if $f \in F_k$, S(f) can be restricted to an interval of unit length by evaluating f at at most N points in D.

Then $L_N(k) = 1$ if $0 \leq N \leq k$, and

$L_N(k) \leq L_{N-p}(k) + L_{N+p-k-2}(k)$ if p is any integer, $1 \leq p \leq (k+1)$.

Proof Let X be an abscissa set in the interval $[0, L_N(k)]$. If X contains fewer than k+1 distinct elements, then the values of f(X), where $f \in F_k$, provide no information about S(f). So

$$L_N(k) = 1 \quad \text{if} \quad N \leq k.$$

On the other hand, S(f) can always be restricted to a proper subinterval of D if X contains (k+1) distinct elements. So, suppose X consists of the points $x_1, x_2, \ldots, x_{k+1}$, where

$$0 \leq x_1 < x_2 < x_3 \ldots < x_{k+1} \leq L_N(k).$$

Let p be an integer, $1 \leq p \leq k+1$. No matter how the strategy proceeds, there must always be a $(k+1)$-abscissa set selected in an interval known to contain S(f), before further restriction of S(f) is possible. So, a further $(k+1)-(p-1)$ points must be selected in the interval $[0, x_p]$, before restriction is possible, after determining that $S(f) \in [0, x_p]$. Thus, if S(f) is to be restricted to a unit interval in the required number of steps, x_p must be chosen such that

$$x_p \leq L_{N-[(k+1)-(p-1)]}(k) .$$

(5.1) i.e. $$x_p \leq L_{N+p-k-2}(k).$$

Similarly, a further p points must be selected in the interval $[x_p, L_N(k)]$ before restriction to a subinterval of this is possible, after determining that $S(f) \in [x_p, L_N(k)]$. So, x_p must be such that

(5.2) $$L_N(k) - x_p \leq L_{N-p}(k).$$

By addition of (5.1) and (5.2),

(5.3) $$L_N(k) \leq L_{N-p}(k) + L_{N+p-k-2}(k).$$

Theorem 5.3 If k is even, and $p = \frac{k+2}{2}$, then $\rho(k) \leq \sqrt[p]{2}$. If k is odd, and $p = \frac{k+1}{2}$, then $\rho(k) \leq \beta_p$, where β_p is the largest positive root of the equation $x^{p+1} = x + 1$.

Proof If k is even, then by equation 5.3,

$$L_N(k) \leq 2\, L_{N-p}(k) .$$

It follows by induction on N, and by the fact that

$L_o(k) = 1$, that

$$L_N(k) \leq (\sqrt[p]{2})^N \quad \text{for all positive integers } N.$$

But then, $\rho_N(k) \leq [(\sqrt[p]{2})^N]^{\frac{1}{N}} = \sqrt[p]{2}$

and so $\rho(k) \leq \sqrt[p]{2}$, if k is even.

If k is odd, and $p = \frac{k+1}{2}$, then by theorem 5.2, $L_N(k) \leq L_{N-p}(k) + L_{N-p-1}(k)$. The inequality $L_N(k) \leq \beta_p^N$ is true if $N = 0$. If it is true for all integers $\leq m$, for some m, then

$$L_{m+1}(k) \leq (\beta_p)^{m+1-p} + (\beta_p)^{m-p}$$

$$= \beta_p{}^{m-p}\,(1 + \beta_p)$$

$$= \beta_p{}^{m+1}.$$

So, by induction, $L_N(k) \leq \beta_p^N$ for all integers $N \geq 0$.

Hence $\rho_N(k) \leq \beta_p$

$\rho(k) \leq \beta_p$, if k is odd.

The theorem below shows that if k is even, then the upper bound $\rho(k) \leq \sqrt[p]{2}$ is best possible. The strategy described is due to C.L. Mallows of Bell Telephone Laboratories, who conjectured that $\rho(k) = \sqrt[p]{2}$ if k is even.

Theorem 5.4 If $k \geq 2$ and k is even, then there is a strategy M for which $\rho(k,M) = \sqrt[p]{2}$, where $p = \frac{k+2}{2}$.

Proof Without loss of generality assume that the domain of F_k is the closed interval $[0, 2(k+2)]$.

Select as the first abscissa set the k+1 points $x_i = 2i$, $i = 1, 2, \ldots, (k+1)$. If $f \in F_k$, and if the appropriate test

function T as defined in lemma 5.1 is evaluated, S(f) can be restricted to either [0, 2(k+1)] or [2, 2(k+2)]. By symmetry, there is no loss of generality in assuming that $S(f) \in [0, 2(k+1)]$.

Now choose x = k+1, that is, the midpoint of [0, 2(k+1)]. There is now a suitable abscissa set in this interval, so, by evaluating the test function T, S(f) can be restricted to either [0, 2k] or [2, 2(k+1)]. Again by symmetry, assume that $S(f) \in [0, 2k]$.

The evaluation of f at x = k-1, allows the restriction of S(f) to an interval of length (2k-2). One successively selects a (k+1)th point in an interval in such a way as to preserve symmetry, and reduces by 2 the length of interval to which S(f) may belong. After p stages, during which f is evaluated at p points in addition to the original abscissa set, S(f) is restricted to an interval of length (k+2), with the values of f known at all points in the interval distant by an integer from the endpoints.

Since this reproduces the original configuration, scaled down by the factor 2, it follows that the procedure may be iterated indefinitely.

Denoting this strategy by M, it follows that if n is an integer of the form $n = (k+1) + \gamma p$, where γ is a positive integer, then

$$d_n(k,M) = \left(\frac{1}{2}\right)^{\gamma} d_o$$

$$\rho_n(k,M) = (2^{\gamma})^{\frac{1}{n}} = 2^{\frac{n-k-1}{np}} .$$

It is now easy to deduce that

$$\lim_{n \to \infty} \rho_n(k,M) = 2^{\frac{1}{p}}$$

i.e. $$\rho(k,M) = 2^{\frac{1}{p}} .$$

Remark In chapter 3, it was shown that the golden mean strategy, though optimal, is not the best procedure if the number N of observations allowed is given in advance of computations. By contrast, the strategy given in theorem 5.4 is also optimal for some finite integers N. For example, take k = 2, p = 2, and N=7. Then, by equation 5.3, $L_7(2) \leq 4$. The strategy M defined in theorem 5.4, gives $d_7(2,M) = \frac{1}{4} d_o$, and so M is optimal for N=7.

For a fuller discussion, we refer to Kiefer [6].

5 - 2 Derivatives of odd order

Known results for the case when k is odd are much less satisfactory, with the one exception k = 1, where $\rho(1) = \beta_1 = \frac{1 + \sqrt{5}}{2}$. I conjecture that the upper bounds $\rho(k) \leq \beta_p$ are best possible, in view of the fact that inequalities in theorem 5.2 are good ones if k is even.

A lower bound for $\rho(k)$ may be obtained by noticing that one may always employ a strategy appropriate for the case k+1. It follows that $\rho(k+1) \leq \rho(k)$ for all integers k, and hence by theorem 5.4, $\rho(k) \geq 2^{\frac{2}{k+3}}$ for all odd k. If $k \geq 3$, $2^{\frac{2}{k+3}} > 2^{\frac{1}{k+1}}$, so this new lower bound is better than that given in lemma 5.1.

Another lower bound may be obtained by using what might be called a multiple version of the golden-mean strategy. This

means simply that for some $\epsilon > 0$, $\frac{k+1}{2}$ distinct points are selected within distance ϵ of each point indicated by the golden-mean strategy. The ideal case $\epsilon = 0$, would allow one to restrict $S(f)$, for $f \in F_k$ from the unit interval D, to an interval of length $\left(\frac{\sqrt{5} - 1}{2} \right)^{n-1}$ by evaluating f at $N = n(\frac{k+1}{2})$ points in D, n being an arbitrary positive integer. This leads to the inequality

$$\rho(k) \geq \left(\frac{\sqrt{5} + 1}{2} \right)^{\frac{2}{k+1}} \quad \text{if } k \text{ is odd.}$$

This inequality is inferior to the bound $\rho(k) \geq 2^{\frac{2}{k+3}}$ except when $k = 1$ or $k = 3$.

In practice, ϵ must be chosen before any computations are made, and since the ratio of ϵ to the length of interval to which $S(f)$ is confined increases steadily, the procedure breaks down unless the number N of evaluations of f is specified in advance. If N is given, ϵ can be chosen small enough such that $\rho_N(k,M)$ is arbitrarily close to

$$\left(\frac{\sqrt{5} + 1}{2} \right)^{\frac{2}{k+3}} .$$

The big difficulty with k odd, $k \geq 3$, lies in finding a strategy such that only one new point is selected at each stage after the first, and such that a certain amount of symmetry is preserved. It is of course possible that the best strategies are not symmetric, but this seems unlikely in view of the results for k even.

We conclude with a brief table giving the best known numerical bounds for $\rho(k)$, for $0 \leq k \leq 6$.

k	lower bound for $\rho(k)$	upper bound for $\rho(k)$
0	2	2
1	$\beta_1 = \frac{\sqrt{5}+1}{2} = 1.618$	$\frac{\sqrt{5}+1}{2} = 1.618$
2	$\sqrt{2} = 1.414$	$\sqrt{2} = 1.414$
3	$\sqrt{\beta_1} = 1.272$	$\beta_2 = 1.325$
4	$\sqrt[3]{2} = 1.260$	$\sqrt[3]{2} = 1.260$
5	$\sqrt[4]{2} = 1.189$	$\beta_3 = 1.207$
6	$\sqrt[4]{2} = 1.189$	$\sqrt[4]{2} = 1.189$

BIBLIOGRAPHY

1. R. Bellman and S. Dreyfus, Applied Dynamic Programming, Princeton University Press, Princeton, N.Y., 1962.

2. A.M. Gleason, Finding the Maximum of a Continuous Function, Proceedings of a Harvard Symposium of Digital Computers and their applications, 3-6 April 1961, Harvard University Press, Cambridge, Massachusetts, 1962, pp. 198-202.

3. O Gross and S. Johnson, Sequential minimax search for a zero of a convex function, Mathematical Tables and other Aids to Computation, Volume 13, 1959, pp. 44-51.

4. T. Hildebrandt, Introduction to the Theory of Integration, Academic Press, New York, 1963.

5. J. Kiefer, Sequential Minimax Search for a Maximum, Proceedings of the American Mathematical Society, Volume 4, 1953, pp. 502-506.

6. _____, Optimal Sequential Search and Approximation Methods under Minimum Regularity Assumptions, Journal of the Society of Industrial and Applied Mathematics, Volume 5, 1957, pp. 105-136.

7. R.S. Pinkham, Random Root Location, Journal of the Society for Industrial and Applied Mathematics, Volume 12, No. 4, December 1964, pp. 855-864.

8. H. Robbins and S. Monro, A Stochastic Approximation Method, Annals of Mathematical Statistics, Volume 22, 1951, pp. 400-407.

9. B. van der Waerden, Modern Algebra, Revised English Edition, Translation by F. Blum, Volume 1, Ungar, New York, 1953.

10. I.M. Yaglom and V. Boltyanskii, Convex Figures, English Translation by P.J. Kelley and L.F. Walton, Holt Rinehart and Winston, New York, 1961.