COMPLEX INTERVAL ARITHMETIC WITH SOME APPLICATIONS

Ray Boche

4-22-66-1 February 1966

LOCKHEED MISSILES & SPACE COMPANY
Lockheed Sunnyvale
Sunnyvale, California

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Interval arithmetic is a part of that branch of numerical analysis concerned with estimating numerical error. It is a means of determining a complete error bound for the errors that occur in digital computation, because of inaccuracies in physical measurements, round-off or truncation of numbers, or truncation of infinite processes. By an error bound we mean an estimate which is correct with probability one, i.e., with certainty.

The form of the error bound will be a number pair representing the closed interval of real numbers for which that number pair serves as end points. For example, [1,2] will represent all real numbers, x, such that $1 \leq x \leq 2$. Such a number pair will be called an interval number. An algebraic system with interval number elements called interval arithmetic is developed as an extension of arithmetic with real numbers in Chapter II. We shall be able to compute with interval numbers in much the same manner as we compute with approximations to real numbers. If we exercise the usual care in computation, we should arrive at a result interval which is a useful error bound.

Interval arithmetic was first suggested by P. S. Dwyer [1] in 1951. Development of interval arithmetic as a formal system and evidence of its value as a computational device was provided by R. E. Moore [2], [3] in 1959 and 1962. Recently Moore and others [4], [5], [6] have developed applications to differential equations. It is also used in matrix computations by Dwyer [7] and E. R. Hansen [8]. Work on the programming of interval arithmetic for digital

computers is included by Moore and others [4], R. E. Boche [9], [10], [11], and S. Shayer [12].

In Chapter II we present the definitions and principal known properties of interval arithmetic and establish some notation. Proofs of several theorems and results in this chapter can be found in Shayer [12]. In the last section of Chapter II two special definitions are included: the first is from [3] and [9]; the second is new, at least formally.

In Chapter III a containment result known to Moore [3], [5], is developed as a major theorem. Chapter III also presents an extension of interval arithmetic into the complex plane along lines suggested by K. Knopp [13].

In Chapter IV some applications known to Moore [5] are presented in a much simplified manner through the containment theorem of Chapter III: in particular, a means of accounting for truncation error in finite approximations to infinite processes.

In Chapter V a computing algorithm for polynomial roots is developed. The algorithm was suggested to the author in outline by E. R. Hansen, and the root counting procedure was suggested by M. Billik.

Motivating Examples

Let us suppose that we perform some computations or "data reduction," using some test results. Suppose that we carry a number of decimal places in the computation, say four, and suppose that the test data are obtained by reading a gauge or another device which can be read accurately to only three places. If the reading is, say, .123+ where the "+" denotes something greater than .123 but less than .124, then the uncertainty in the reading could be represented by

computing with the interval number [.1230, .1240]. If the computation is long and complex, it may be impossible to determine by a priori analysis the effect of computing with an inexact number; however, the interval result of a corresponding interval computation automatically takes into account any effects of such inaccuracies.

Although uncertainty in physical measurement is a well-known phenomenon, it is only a minor justification for interval arithmetic. A far more important (though lesser known) source of uncertainty in digital computation is round-off error. The effects of round-off error were seldom encountered in any dramatic fashion prior to the advent of modern high speed computers.

A digital computer represents all real numbers as rational numbers of finite length. Therefore, in computing with an irrational number, there will certainly be round-off error in the rational approximation. Similarly, if the representation of a rational number requires more digits than the "word length" of the computer at hand, round-off error will result. Also those rational numbers which are repeating decimal numbers such as 1/3 will obviously produce round-off error. In order to bound round-off error we may represent Pi by the interval number [3.1415926, 3.1415927] and 1/3 by [.33333333, .33333334]. Additionally we must consider the fact that most computers use the binary number system. The decimal number .1, for example, is equal to the repeating binary number $.0001\overline{1001}$, and again a round-off error results.

In computations which involve relatively few operations round-off error may have little effect on the accuracy of the result, but round-off error becomes an important concern in computations involving thousands or even millions of arithmetic operations.

Computation 1. Suppose we are to compute sine $n\pi$, $n = 0,1,2,\ldots$. A typical digital computer program computes sine x where $-\pi/2 < x \leq \pi/2$ accurately by means of an approximating function valid in the specified range. For values of the argument outside the specified range the argument is "scaled," a process which in this case consists of dividing by π and using a remainder as argument. As n increases, dividing by an approximation to Pi, remultiplying, and subtracting will almost certainly result in round-off error and, in turn, a non-zero result.

Computation 2. Suppose we are to compute the value of the determinant

$$\begin{vmatrix} 3 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{vmatrix}.$$

The steps in the reduction will be indicated using four-decimal places with rounding by truncation. The method used is a reduction to triangular form using row operations. Dividing by 3 and subtracting gives

$$\begin{vmatrix} 3 & 2 & 1 \\ 0 & .3334 & .6667 \\ 0 & 1 & 2 \end{vmatrix}.$$

Interchanging Rows 2 and 3, dividing by 3, and subtracting gives

$$-\begin{vmatrix} 3 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & .0001 & .0001 \end{vmatrix},$$

from which the program concludes that the value of the determinant is -.0003.

This computation was encountered when a large program run on the IBM 7094 computer failed. The inversion of a singular matrix caused the failure. A singular matrix may be recognized by most programs only by detecting

a zero determinant. In the case at hand both single and double precision arithmetics give similar incorrect results. Although the example here depends on the logic of the computer's arithmetic unit, it is safe to assume that for any given logical circuitry a similar example could be constructed.

Computation 3. Symmetric rounding is no assurance of accuracy. For instance, $(2/3 - 1/3 - 1/3) \cdot 9000 \cdot 9000$ should equal zero. However, with

.6667	.3334	.0001		.9000
-.3333	-.3333	×9000.		×9000.
.3334 ,	.0001 ,	.9000 ,	and	8100.

we see that the slight difference between zero and non-zero can be magnified considerably. This computation also indicates another problem related to rounding. This problem of subtracting out significance is called cancellation. The input numbers indicate four significant figures. However, after the second subtraction the result contains, at most, one significant figure.

Interval arithmetic does not prevent round-off error; it bounds all the round-off error that may have occurred. Thus while an ordinary computation provides a result of completely unknown accuracy, interval arithmetic provides simultaneously, information on the accuracy of a result.

CHAPTER II

INTERVAL ARITHMETIC

Definition 2.1: A set consisting of a closed interval of real numbers, x, such that $a \leq x \leq b$ is called an interval number.

An interval number will be denoted by an upper case letter. When it is desirable to emphasize that operations on the closed intervals may be accomplished by operating on the end points, we use the notation $[a, b]$. In other instances, a standard set notation is helpful; thus the following are equivalent: $A = [a, b] = \{x | a \leq x \leq b\}$.

Two interval numbers $[a, b]$ and $[c, d]$ are equal if and only if $a = c$ and $b = d$. Equality of interval numbers is an equivalence relation.

The four basic arithmetic operations for interval numbers are defined below. In adding two interval numbers $[a, b]$ and $[c, d]$, since $a \leq x \leq b$ and $c \leq y \leq d$, the smallest possible value in the set of sum $x + y$ is $a + c$, and the largest is $b + d$. This concept motivates the definition of interval arithmetic.

Definition 2.2: $[a, b] + [c, d] = [a + c, b + d]$.

Definition 2.3: $[a, b] - [c, d] = [a - d, b - c]$.

Definition 2.4: $[a, b] \cdot [c, d] = [\min(ac, ad, bc, bd), \max(ac, ad, bc, bd)]$.

Definition 2.5: If $0 \notin [c, d]$,

$$[a, b]/[c, d] = [\min(a/c, a/d, b/c, b/d), \max(a/c, a/d, b/c, b/d)].$$

The interval numbers $[0, 0]$ and $[1, 1]$ serve as additive and multiplicative identities, respectively. Interval arithmetic is closed, associative, and commutative with respect to addition and multiplication (See Shayer [12]).

Definition 2.6: The width of the interval number $[a,b]$ is $b-a$.

A real number can be thought of as corresponding to an interval number of width zero. The correspondence is indicated by $a \longleftrightarrow [a,a]$, and considered as an embedding of the real numbers in the interval numbers where they appear as interval numbers of zero width.

Elementary Theorems, Properties, and Further Definitions

Theorem 2.1: Additive inverses exist only for interval numbers of zero width.

Theorem 2.2: Multiplicative inverses exist only for non-zero interval numbers of zero width.

Theorem 2.3: If $A + B = A + C$, the cancellation law for addition holds, and $B = C$.

Theorem 2.4: If $AB = AC$ and $0 \notin A$, the cancellation law for multiplication holds, and $B = C$.

Proof: Let $A = [a,b]$, $B = [c,d]$, and $C = [e,f]$. Consider the case $0 < a \leq b, 0 \leq c \leq d$. We have $AB = AC = [ac,bd]$. The $\min(ae, af, be, bf)$ must equal ac. As $0 < a$ and $0 \leq c$, then $0 \leq ac$, and hence $0 \leq e \leq f$. Therefore the minimum must be ae or be. But $a \leq b$ and $0 \leq e$; so the minimum is ae. Since $ac = ae$, $c = e$. Similarly we may show that $d = f$, and therefore $B = C$.

The case $0 < a \leq b, c \leq 0 \leq d$ may be treated in the same manner. Multiplication by $[-1,-1]$ allows the other possible cases to be included in the two above, and we conclude that the cancellation law for multiplication holds.

Theorem 2.5: The distributive law for interval numbers does not hold in general.

Proof: By counter example, $[0,1] \cdot ([1,2] + [-1,0]) = [0,1] \cdot [0,2] = [0,2]$, but $[0,1] \cdot [1,2] + [0,1] \cdot [-1,0] = [0,2] + [-1,0] = [-1,2]$.

Interval numbers may be partially ordered by the relations of set inclusion and less than:

$[a,b] < [c,d]$ if and only if $b < c$,

$[a,b] \subset [c,d]$ if and only if $c \leq a \leq b \leq d$.

Theorem 2.6: The following relation called subdistributivity holds for interval numbers: $A(B + C) \subset AB + AC$.

Theorem 2.7: In the special case where the interval number A is of zero width, $A(B + C) = AB + AC$.

Definition 2.7: The union of two interval numbers, $A \cup B$, is the interval number $\{x \mid x \in A \text{ or } x \in B\}$.

Definition 2.8: The intersection of two interval numbers, $A \cap B$, is the interval number $\{x \mid x \in A \text{ and } x \in B\}$.

Finally we note that the set of interval numbers forms an Abelian semigroup under the operation of addition and also under the operation of multiplication.

Illustrative Computations

Let $A = [4,5]$, $B = [-3,-2]$, and $C = [0,1]$. Then we may perform the computation $AX^2 + BX + C$ as follows.

First let $X = [1,2]$. Then $X^2 = [1,2] \cdot [1,2] = [1,4]$ by Definition 2.4. Also by Definition 2.4, $AX^2 = [4,5] \cdot [1,4] = [4,20]$, and $BX = [-3,-2] \cdot [1,2] = [-6,-2]$. By Definition 2.2, $AX^2 + BX + C = [4,20] + [-6,-2] + [0,1] = [-2,19]$.

If the expression $(AX + B)X + C$ is evaluated with $A, B, C,$ and X as above, the result is different: $AX + B = [4,5] \cdot [1,2] + [-3,-2] = [1,8]$ by

Definitions 2.4 and 2.2; $(AX + B)X + C = [1,8] \cdot [1,2] + [0,1] = [1,17]$ also by Definitions 2.4 and 2.2.

We note that although the real expressions $ax^2 + bx + c$ and $(ax + b)x + c$ are equivalent, the corresponding interval expressions are not. In fact $(AX + B)X + C \subset AX^2 + BX + C$ as it must be according to Theorem 2.6. In this example the containment is proper containment and we have the relation $[1,17] \subset [-2,19]$.

The widths of the interval numbers $[1,17]$ and $[-2,19]$ are 16 and 21 , respectively, by Definition 2.6. According to Definitions 2.7 and 2.8, their union is $[-2,19]$ and their intersection is $[1,17]$.

Finally we note that computations with interval numbers do not always grow in width with additional operations. As an illustration, consider the recursive computation $X_{n+1} = X_n \cdot \left[\frac{1}{4}, \frac{1}{2}\right]$, and let $X_o = [0,1]$. Then $X_n = [0, 1/2^n]$, and we see that as n increases, the width of the result interval decreases.

Two Special Definitions

According to our definition of multiplication, if $A = [-1,1]$ and $B = [-1,1]$, then $AB = [-1,1]$. This result is acceptable providing A and B are coincidentally equal. However, if this same result were obtained as $A^2 = [-1,1]$, we would be concerned about admitting negative numbers into a set of squares of real numbers. This problem can be alleviated by making careful note of the identity of factors in the product A^2 and defining the resultant set accordingly.

Definition 2.9: $A^2 = \{x^2 | x \in A\}$.

In contrast, the product $A \cdot A = \{xy | x \in A,\ y \in A\}$.

If A and B are as above, then $A - B = [-2, 2]$. This expression could be written as $A + (-B) = [-2, 2]$ where $-B$ implies $0 - B$. Again this result is acceptable if the equality of A and B is coincidental. However, if we wish to represent by $A - A$ the subtracting of A from itself, we have identity rather than equality. The definition for subtraction gives

$A - A = \{x - y | x \in A, y \in A\} \neq [0, 0]$ in general.

Definition 2.10: The sum of an interval number and its negative is $[0, 0]$.

In this last definition the possessive form "its negative" is used to imply identity, and symbolically $A + (-A) = \{x - x | x \in A\} = [0, 0] = 0$.

CHAPTER III

COMPLEX INTERVAL ARITHMETIC

Extension to the Complex Plane

There is no need to limit the application of interval numbers to the measure of uncertainty in real numbers. We would like to use interval numbers to determine a region of uncertainty in computing with complex numbers. No new difficulties arise if we choose the Cartesian representation for complex numbers.

As is customary, we denote the complex numbers as an ordered pair of real numbers, (a, a') .

Definition 3.1: The sum of two complex numbers, (a, a') and (b, b') , is the complex number $(a + b , a' + b')$.

Definition 3.2: The product of two complex numbers, (a, a') and (b, b') , is the complex number $(ab - a'b' , ab' + a'b)$.

Under these definitions we note that the real number, a , corresponds to and may be identified with the complex number, $(a, 0)$. Traditionally the complex number, $(0, 1)$ is denoted by the letter i . Then since $(0, 1) \cdot (0, 1) = (-1, 0)$, $i^2 = -1$. In summary the complex number (a, a') , could be represented as

$$(a, a') = (a, 0) + (0, a') = a + a'(0, 1) = a + a'i .$$

We wish to develop complex interval numbers in a similar manner.

Definition 3.3: A complex interval number is an ordered pair of interval numbers (A, B) .

Upper case script letters are used to denote complex interval numbers. As in the case with ordinary complex numbers, we shall have occasion to refer to the complex interval number, $([0, 0], [1, 1])$ as "i."

In set notation a complex interval number may be represented in the form $\mathcal{A} = [a,b] + [c,d]\,i = \{x + yi \mid a \leq x \leq b,\ c \leq y \leq d\}$. Geometrically a complex interval number may be conceived of as a closed rectangular region in the complex plane (see Figure 3.1). In terms of vectors we may think of the set of all vectors whose initial point is the origin and whose terminal point lies in or on the boundary of the rectangular region.

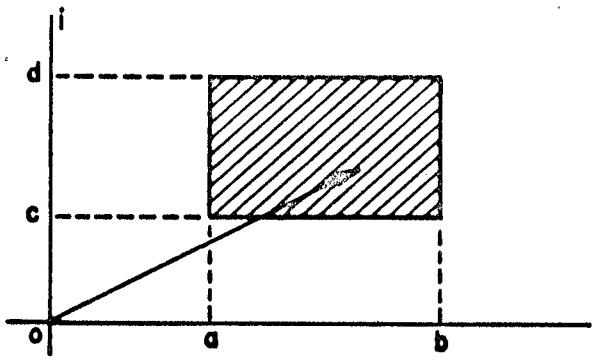

Figure 3.1

If we attempt to express the set of complex numbers represented by a complex interval number, $\mathcal{A}$, in its polar form, we encounter certain difficulties. Although the set of complex numbers determined by the absolute value and the amplitude of $\mathcal{A}$ contains all the complex number elements of the complex interval number $\mathcal{A}$, it will in general also contain a great many others. The shaded portions of Figure 3.2 represent additional complex numbers included in a polar representation of a complex interval number.

Such a polar representation could serve as the basis for an alternative definition of complex interval numbers. We could of course determine a rectangular region containing all elements of such a polar representation. However, it

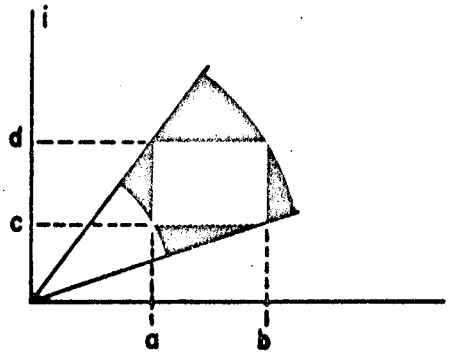

Figure 3.2

is easy to see that the only complex interval numbers for which the polar and Cartesian representations are equal have one of three forms:

$$[a,a] + [b,b]\ i\ ,$$

$$[c,d] + [0,0]\ i\ ,$$

or

$$[0,0] + [c,d]\ i\ ,$$

where 0 is not contained in the open interval (c,d).

Definition 3.4: The conjugate of a complex interval number, $\mathcal{A}$, is that complex interval number, $\mathcal{B}$, which determines a region symmetric to $\mathcal{A}$ with respect to the axis of reals.

Definition 3.5: The negative of a complex interval number, $\mathcal{A}$, is that complex interval number, $\mathcal{B}$, which determines a region symmetric to $\mathcal{A}$ with respect to the origin ($\mathcal{B}$ may be denoted $-\mathcal{A}$).

Definition 3.6: The sum of two complex interval numbers (A, A') and (B, B') is the complex interval number $(A + B,\ A' + B')$. With the alternative notation,

$$([a,b] + [c,d]\ i) + ([e,f] + [g,h]\ i) = [a + e,\ b + f] + [c + g,\ d + h]\ i\ .$$

Definition 3.7: The product of two complex interval numbers (A, A') and (B, B') is the complex interval number $(AB - A'B', AB' + A'B)$.

Unfortunately the properties of ordinary complex conjugates do not hold in general for complex interval conjugates, because of the lack of additive inverses in interval arithmetic. For example, by Definition 3.1 the sum of the conjugate complex numbers $(a, b) + (a, -b) = (2a, 0)$ is equal to $2a$, a real number. However, by Definition 3.6 the sum of the conjugate complex interval numbers $(A, B) + (A, -B) = (2A, B-B)$ is not equal to $2A$ since $B - B = [a, b] - [a, b] = [a-b, b-a] \neq [0, 0]$ unless $a = b$. Consider the product of the conjugate complex interval numbers $(A, B) \cdot (A, -B) = (A^2 + B^2, AB - AB) \neq (A^2 - B^2, 0)$, and again we see that the failure of additive inverses causes the sacrifice of another important property of complex numbers.

The difficulties arising from the lack of additive inverses may be avoided in some instances. In considering the product of a complex interval number and its conjugate above, we indicated our recognition of identity of interval numbers by denoting the product $A \cdot A$ as A^2. This notation is in accordance with Definition 2.9 and serves to insure that negative elements are not included in a set of self products of real numbers. If care is taken in the distinction of identity, the product of a complex interval number and its conjugate may be defined in such a way that the product will be a real interval number. The phrase, "its conjugate," is used below to imply identity except for sign in the second component of the ordered pairs of interval numbers comprising complex interval numbers. Specifically Definitions 3.8 and 3.9 below do not apply to the complex interval numbers (A, B) and $(A, -C)$ even if $B = C$.

Definition 3.8: The sum of a complex interval number (A, B) and its conjugate $(A, -B)$ is the complex interval number $(2A, 0)$.

Definition 3.9: The product of a complex interval number (A, B) and its conjugate $(A, -B)$ is the complex interval number $(A^2 - B^2, 0)$.

Using the phrase "its negative" in a similar manner, we have the following definition, corresponding to Definition 2.10, for real interval numbers.

Definition 3.10: The sum of a complex interval number (A, B) and its negative $(-A, -B)$ is the complex interval number $(0, 0)$.

The quotient of two complex interval numbers may now be defined.

Definition 3.11: If (A, A') and (B, B') are two complex interval numbers and $0 \notin (B, B')$, their quotient, $(A, A')/(B, B')$, is defined as the complex interval number (C, C') where

$$(C, C') = \frac{(A, A')}{(B, B')} = \frac{(A, A') \cdot (B, -B')}{(B, B') \cdot (B, -B')} = \frac{(A, A') \cdot (B, -B')}{(B^2 - B'^2, 0)}$$

$$= \frac{(AB + A'B', A'B - AB')}{(B^2 - B'^2, 0)} = \left(\frac{AB + A'B'}{B^2 - B'^2}, \frac{A'B - AB'}{B^2 - B'^2} \right)$$

providing $0 \notin B^2 - B'^2$. Intermediate steps in the definition show that it is consistent with previous definitions.

In addition to Definitions 3.7 and 3.9 we have employed a correspondence between complex interval numbers and real interval numbers analogous to the similar correspondence between complex numbers and real numbers, i.e., $A \longleftrightarrow (A, 0)$. We have also ascribed to complex interval numbers another analog to a property of complex numbers; namely, $a \cdot (b, c) = (ab, ac)$, or for intervals, $A \cdot (B, C) = (AB, AC)$. This later property is also a consequence of the correspondence just noted. We may employ other properties of a similarly analogous

nature without specific elaboration. However, those properties of complex interval numbers developed here should prove adequate for the present work.

Finally we note that if we are to employ complex interval numbers in computation successfully, special care must be taken to develop conjugate pairs simultaneously and consciously whenever possible.

Containment Theorem

The following four lemmas will be needed for the containment theorem to follow.

Let I, J, K, and L be interval numbers such that $I \subset K$ and $J \subset L$.

Lemma 3.1: $I + J \subset K + L$.

Proof: Let $I = [a,b]$, $J = [c,d]$, $K = [e,f]$, and $L = [g,h]$. Then from the definition of interval addition, $I + J = [a + c, b + d]$, and $K + L = [e + g, f + h]$. Since $I \subset K$ and $J \subset L$, $e \leq a$, $g \leq c$, $b \leq f$, and $d \leq h$. Combining these relations gives $e + g \leq a + c$ and $b + d \leq f + h$. Hence $I + J \subset K + L$.

Lemma 3.2: $I - J \subset K - L$.

Proof: Let I, J, K, and L be as in Lemma 3.1. Then from the definition of interval subtraction, $I - J = [a-d, b-c]$, and $K - L = [e-h, f-g]$. Combining the relations $e \leq a$ and $d \leq h$ gives $e - h \leq a - d$. Similarly, from $b \leq f$ and $g \leq c$ we get $b - c \leq f - g$. Hence $I - J \subset K - L$.

Lemma 3.3: $I \cdot J \subset K \cdot L$.

Proof: With the set notation, $I \cdot J = \{w \cdot x \mid a \leq w \leq b, c \leq x \leq d\}$, and $K \cdot L = \{y \cdot z \mid e \leq y \leq f, g \leq z \leq h\}$. Then $e \leq a \leq b \leq f$ and $g \leq c \leq d \leq h$, and we note that every element, $w \cdot x$, of the set, $I \cdot J$, is also an element, $y \cdot z$, of the set $K \cdot L$. Therefore $I \cdot J \subset K \cdot L$.

Lemma 3.4: $I/J \subset K/L$ if $0 \notin L$.

Proof: Again, with the set notation, $I/J = \{w/x \mid a \leq w \leq b, c \leq x \leq d\}$, and $K/L = \{y/z \mid e \leq y \leq f, g \leq z \leq h\}$. Then, using $e \leq a \leq b \leq f$ and $g \leq c \leq d \leq h$, we find that every element, w/x, of the set I/J is also an element, y/z, of the set K/L. Therefore, if $0 \notin L$, $I/J \subset K/L$.

Theorem 3.1: Let $F(X_1, X_2, \ldots, X_n)$ be a rational expression in the interval variables $X_1, X_2, \ldots, X_n$. If $Y_1 \subset X_1, Y_2 \subset X_2, \ldots, Y_n \subset X_n$, then $F(Y_1, Y_2, \ldots, Y_n) \subset F(X_1, X_2, \ldots, X_n)$.

Proof: Every interval arithmetic operation specified by $F(Y_1, Y_2, \ldots, Y_n)$ may be associated with an interval arithmetic operation in $F(X_1, X_2, \ldots, X_n)$.

Let $\otimes$ specify any of the four interval arithmetic operations: addition, subtraction, multiplication, or division. If I and J each represent either an interval constant or one of the interval variables Y_i and if K and L each represent either the corresponding interval constant or the corresponding interval variable X_i, then $I \subset K$ and $J \subset L$. Then as each interval operation is performed, $I \otimes J \subset K \otimes L$ by the appropriate lemma, 3.1, 3.2, 3.3, or 3.4; we now have a reduced expression in which each interval constant, interval variable, or interval valued sub-expression appearing in $F(Y_1, Y_2, \ldots, Y_n)$ is contained in the corresponding interval constant, interval variable, or interval valued sub-expression appearing in $F(X_1, X_2, \ldots, X_n)$. Since the number of operations is finite, we conclude that $F(Y_1, Y_2, \ldots, Y_n) \subset F(X_1, X_2, \ldots, X_n)$.

A particularly important application of Theorem 3.1 is the case in which the intervals Y_i and the constant intervals employed are all of width zero. Then by the correspondence exhibited in Chapter II we may conclude that if $y_1 \in X_1, y_2 \in X_2, \ldots, y_n \in X_n$, then $F(y_1, y_2, \ldots, y_n)$ is a real number, and

further the function of real values $f(y_1, y_2, \ldots, y_n) \in F(X_1, X_2, \ldots, X_n)$. Conversely if $f(y_1, y_2, \ldots, y_n)$ is a real valued function of the real values $y_1, y_2, \ldots, y_n$ and if each $y_1 \in X_1, y_2 \in X_2, \ldots, y_n \in X_n$ and if each operation specified by f is defined for the interval variables X_i, we associate the functions $f(y_1, y_2, \ldots, y_n)$ and $F(X_1, X_2, \ldots, X_n)$ and refer to F as an interval valued function.

Definition 3.12: Given a real valued function, $f(x_1, x_2, \ldots, x_n)$, where each of the real variables, x_i, may assume any of the set of real values, X_i, we define the corresponding interval valued function, $F(X_1, \ldots, X_n)$, by the set $F(X_1, X_2, \ldots, X_n) = \{y \mid y = f(x_1, \ldots, x_n),\ x_i \in X_i\ (i = 1, \ldots, n)\}$.

Containment Theorem in the Complex Plane

The four basic arithmetic operations have now been defined for complex interval numbers. If we consider any one of the four operations denoted by $\otimes$, we have $\mathcal{A} = \mathcal{B} \otimes \mathcal{C}$ or $(A, A') = (B, B') \otimes (C, C')$ where each of the interval numbers A and A' is some interval valued function of the interval variables B, B', C, C'. We may say $A = F_1(B, B', C, C')$ and $A' = F_2(B, B', C, C')$. Now suppose we have complex interval numbers $\mathcal{I}$, $\mathcal{J}$, $\mathcal{K}$, and $\mathcal{L}$ such that $\mathcal{I} \subset \mathcal{K}$ and $\mathcal{J} \subset \mathcal{L}$. In expanded notation, $I \subset K$, $I' \subset K'$, $J \subset L$, and $J' \subset L'$. Again if $\otimes$ represents any of the four arithmetic operations where they are defined, we have the following theorem for complex interval numbers.

Theorem 3.2: $\mathcal{I} \otimes \mathcal{J} \subset \mathcal{K} \otimes \mathcal{L}$.

Proof: Let $\mathcal{I} \otimes \mathcal{J} = (Y, Y')$ where $Y = F_1(I, I', J, J')$ and $Y' = F_2(I, I', J, J')$. Let $\mathcal{K} \otimes \mathcal{L} = (X, X')$ where $X = F_1(K, K', L, L')$ and $X' = F_2(K, K', L, L')$. Then by Theorem 3.1, $Y \subset X$ and $Y' \subset X'$. Therefore $\mathcal{I} \otimes \mathcal{J} \subset \mathcal{K} \otimes \mathcal{L}$.

The extension to a general containment theorem for rational expressions of complex interval numbers and a formal definition of complex interval functions may now be made to parallel the development above for interval numbers and interval functions. Again, in the special case where a function of complex interval numbers consists only of single complex numbers, each in turn contained in some complex interval number, the function of complex variables may be associated with the function of complex interval variables, and in all cases the resulting complex number will be an element of the corresponding complex interval number.

CHAPTER IV

APPLICATIONS

The procedures described in this chapter are intended to illustrate the broad applicability of interval arithmetic. Each procedure could be developed into an effective computing algorithm. The guaranteed error bounding makes each of them of interest as a diagnostic tool, even in the present form.

Summation of Series

Let $s_n(x) = \sum_{i=0}^{n} a_i(x)$. If $a_i(x) = f(x,i)$ and $A_i(X) = F(X,i)$, then Theorem 3.1 includes $S_n = \sum_{i=0}^{n} A_i(x)$, and $s_n \in S_n$. Further, if s_n is the n-th partial sum of an infinite series, uniformly convergent for $x \in [a,b]$, $s(x) = \sum_{i=0}^{\infty} a_i(x)$; and if $s(x) = s_n(x) + r_n$ where r_n is the remainder term which satisfies

$$|r_n| = \left| \sum_{i=n+1}^{\infty} a_i(x) \right| < \delta(n)$$

for n sufficiently large and for all $x \in [a,b]$, then letting $R_n = [-\delta(n), \delta(n)]$, $S(X) = S_n(X) + R_n$ and $s(x) \in S(X)$. Thus interval arithmetic can be extended directly from rational interval functions to those elementary functions for which a series representation of the indicated form exists.

Evaluation of Definite Integrals

Let $f(x)$ be a real valued function such that $\int_a^b f(x)\,dx$ exists and such that the associated interval valued function $F(X)$ is defined for $X = [a,b]$. Now let $Y = [c,d] = F(X)$, and let $y = f(x)$. If x takes on any value such that $a \le x \le b$, we observe directly from Theorem 3.1 that $y \in Y$, and hence

we may conclude that $Y = [c,d]$ includes all values that $f(x)$ may take on in the interval $[a,b]$. In particular we note that c is a lower bound and d is an upper bound for f in $[a,b]$.

From Figure 4 we see that $d(b-a)$ and $c(b-a)$ are an upper and lower bound, respectively, on $\int_a^b f(x)\,dx$. Selecting intermediate points $a < a_1 < a_2 < \ldots < a_n < b$ leads to a computing algorithm since

$$\int_a^b f(x)\,dx = \int_a^{a_1} f(x)\,dx + \int_{a_1}^{a_2} f(x)\,dx + \ldots + \int_{a_n}^b f(x)\,dx .$$

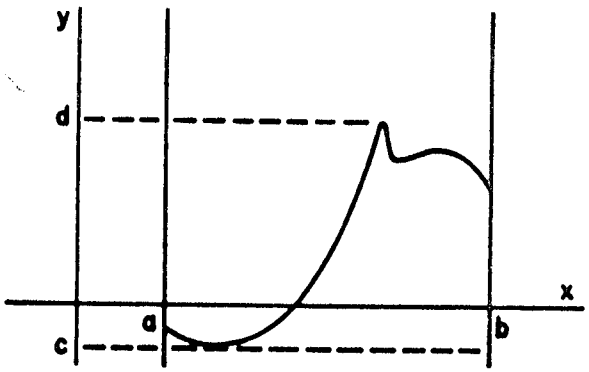

Figure 4

Isolation of Roots

Let $f(x)$ and $F(X)$ be as in the previous section. Then if the interval valued function $F(X)$ is evaluated over the interval $[a,b]$, we know that in the result interval, $Y = [c,d]$, c is a lower bound and d is an upper bound on $f(x)$ over $[a,b]$. Therefore if $f(x)$ has a real root in the interval $[a,b]$, the interval $[c,d]$ must contain zero. If $0 \in [c,d]$, subdivision of $[a,b]$ and re-evaluation may continue until all real roots in the interval have been located, a process again leading to a computing algorithm.

CHAPTER V

A COMPUTING ALGORITHM FOR POLYNOMIAL ROOTS

Let $P(x)$ be an arbitrary polynomial of finite degree with real coefficients, i.e., $P(x) = \sum_{i=0}^{n} a_i x^i$ with the a_i all real. From the fundamental theorem of algebra we know that $P(x)$ may be represented as a product of factors of the form $P(x) = a_n \prod_{i=0}^{n} (x - r_i)$ where the r_i are complex numbers. The relevant problem here is that of determining the roots, r_i, of a polynomial of this type. Several numerical methods for this purpose are well known and in common use but few (if any) known methods can claim to be always successful or even nearly always correct. We propose to develop here a method employing interval arithmetic which will be nearly always successful and, far more important, always correct. The significance of the claim "always correct" becomes clear in considering the problems inherent in implementing known methods on a digital computer. Although not all difficulties are common to all methods, one of those most frequently encountered is the direct result of round-off error that occurs because only numbers of finite length can be used in digital computers. Interval arithmetic allows us to compute with a digital computer in a manner which provides guaranteed error bounds together with results. Thus there is no need to be concerned with such common occurrences as an iterative numerical method converging to an incorrect root or failing to converge because of round-off error.

The most important consideration in recognizing the need for the program proposed is encountered in facing the numerical difficulties entailed in asking whether or not a given complex number is a root of a particular polynomial.

Again, because of round-off error, we cannot answer such a question. Then, as might be anticipated, in general the coefficients of a polynomial of reduced degree cannot be determined with accuracy and confidence even when we believe that a root of the original polynomial has been correctly identified.

In the remainder of this chapter we shall devote our attention to the development of a computing algorithm for polynomial roots. It should be recognized that the selection of particular procedures to be followed, the order in which procedures are to be applied, and the choice of many computing parameters are arbitrary. The goal is to bound the roots of a polynomial as closely as possible and practical within the constraints imposed by the word length of the computer at hand. In particular, multi-precision arithmetics and fixed point arithmetics are specifically excluded from consideration here.

Determining a Region Containing All Roots

Let $P(x) = \sum_{i=0}^{n} a_i x^i$ be an arbitrary polynomial (of finite degree) with real coefficients. Now suppose that r is a root of $P(x)$ and $|r| \geq 1$. Then

$$P(r) = \sum_{i=0}^{n} a_i r^i = 0 ;$$

$$-a_n r^n = \sum_{i=0}^{n-1} a_i r^i .$$

Divide by $a_n r^{n-1}$:

$$-r = \sum_{i=0}^{n-1} \frac{a_i r^i}{a_n r^{n-1}} = \sum_{i=0}^{n-1} \frac{a_i}{a_n r^{n-1-i}} .$$

By using the "triangle inequality,"

$$|r| \leq \sum_{i=0}^{n-1} \left| \frac{a_i}{a_n r^{n-1-i}} \right| ,$$

and since $|r| \geq 1$,

$$|r| \leq \sum_{i=0}^{n-1} \left| \frac{a_i}{a_n} \right| = \frac{1}{|a_n|} \sum_{i=0}^{n-1} |a_i| = s .$$

Therefore $|r| \leq s$ whenever $|r| \geq 1$, and s may be readily computed from the coefficients of the original polynomial.

Suppose instead that $|r| \leq 1$. Then

$$-a_n r^n = \sum_{i=0}^{n-1} a_i r^i .$$

Divide this time by a_n and note that $|r| \leq 1$:

$$|-r^n| \leq \frac{1}{|a_n|} \sum_{i=0}^{n-1} |a_i| = s ,$$

or $|r| \leq s^{1/n}$ whenever $|r| \leq 1$.

Now if t is the larger of s and $s^{1/n}$, all roots of the polynomial $P(x)$ lie in a circle of radius t about the origin. It is convenient to note that the circle of radius t and hence all roots are contained in the single complex interval number $T = [-t,t] + [-t,t]\, i$.

Determining Approximate Roots

The algorithm to be developed in this chapter requires, as a starting point, some approximate roots. The Quotient-Difference algorithm (see Henrici [14]) deserves special consideration. However, any standard technique may be used. If the technique selected also requires some starting approximations,

these may be determined in an arbitrary but systematic manner from within the region containing all roots. This procedure is not as haphazard as it may sound, because the major algorithm will provide a means of both refining any approximation obtained and also determining an error bound.

Interval arithmetic may be used to free the algorithm from many of the constraints imposed by ordinary machine arithmetics. For example, we could select those approximate roots whose error bound satisfies predetermined criteria and use interval arithmetic to determine with error bounds the interval coefficients of a polynomial of reduced degree. Then we could proceed with a presumably simpler problem.

Determining an Error Bound for an Approximate Root

Suppose that z is an approximate root of $P(x)$. Then

$$P(z) = a_n \prod_{k=1}^{n} (z - r_k) ,$$

or

$$\left| \frac{P(z)}{a_n} \right| = \prod_{k=1}^{n} |z - r_k| .$$

Suppose that z is approximately equal to root, r_1, and assume that $|z - r_1| \leq |z - r_k|$ for $(k = 2, \ldots, n)$. Then

$$\left| \frac{P(z)}{a_n} \right| \geq |z - r_1|^n ,$$

and

$$|z - r_1| \leq q = \left| \frac{P(z)}{a_n} \right|^{1/n} .$$

Since $z = y + yi$, a root must lie in the complex interval number or "box" B: $B = [x - q, x + q] + [y - q, y + q]\,i$. Again B contains a circular region which in turn contains a root.

Improving the Error Bound for an Approximate Root

If r is a root of $P(z)$ and z is an approximate root, then $r = z + h$, and $P(z + h) = 0$. By expansion about z,

$$P(z + h) = P(z) + hP'(z) + \frac{h^2}{2!} P''(z) + \ldots \quad .$$

If terms of order h^2 or higher are neglected (since presumably h is small), we obtain $h = -\frac{P(z)}{P'(z)}$. Correcting the approximate root z by the approximate error h yields the well-known Newton's method,

$$z_{n+1} = z_n - \frac{P(z_n)}{P'(z_n)} \, .$$

In the previous section a method for determining an error bound for an approximate root was presented. Denote the box comprising the error bound by B_o, and let z_o be the approximate root used to determine B_o. Then if we apply the Newton iteration n times and apply the method of the previous section to the new approximate root, z_n, we get a new box, B_1, which also contains a root.

If the process fails to converge for a particular value of z_o or converges to some root not in B_o, $B_o \cap B_1$ may be null, and we will elect to terminate the process. The process will also be terminated if $B_o \subset B_1$.

If both B_o and B_1 are "good" error bounds and $B_o \cap B_1$ is non-null, we assume that round-off error may be a factor and that their intersection contains a root. If the center of the box $B_o \cap B_1$ is selected as a new approximate

root, z_o', an error bound, B_o', may be determined about z_o' by the method of the previous section. If (as assumed) B_o and B_1 contain the same root, we would expect B_o' to be smaller than either. If so, we replace B_o with B_o' and z_o with z_o' and begin again; if not, we select the smaller of B_o and B_1 and consider the process terminated.

Counting Roots

Suppose that we have determined a region containing one or more roots. We need to determine the number of roots (not necessarily distinct) in the region. A promising method for counting roots is outlined below. The exact conditions of applicability and a detailed proof are items for further study.

The basis of our contention is two-fold:

1. In progressing around the boundary of a region of the complex plane known to contain one or more roots, the argument of the function progresses through a period of 2π n times where n is the number of enclosed roots.

2. A change of π in the argument is determined when the boundary is intersected by both the "level curves," U_o and V_o.

The example below illustrates the procedure.

Let $P(z) = z^2 + 4 = 0$ and suppose we have determined a box containing a root, $B = [-1,1] + [1,3]\,i$. Evaluation along the right-hand edge shows

$$P(1 + [1,3]\,i) = ([1,1] + [1,3]\,i)^2 + [4,4]$$

$$= ([1,1] - [1,9] + 2[1,3]\,i) + [4,4] = [-4,4] + [2,6]\,i = U + Vi\,.$$

Since $0 \in U$ but $0 \notin V$, a useful piece of information has been obtained, i.e., the argument of $P(z)$ has not changed by as much as π. Evaluating along the top gives $U + Vi = [-5,-4] + [-6,6]\,i$, and since $0 \notin U$ and $0 \in V$, the

argument has now changed by π . A similar evaluation along the left-hand edge and the bottom again shows the argument has changed by π ; we conclude that the region B contains exactly one root of P(z) .

An Algorithm

The algorithm described here is of course, not intended for direct implementation. Rather it is intended to illustrate how the sections of this chapter might fit together to satisfy our objective.

The algorithm as stated contains abundant opportunities for failure. However, even in this crude form it is "always correct" if properly programmed. That is, it detects its own failures.

The procedures described in this chapter should prove adequate for the development of a considerably refined and nearly always successful algorithm. Although a complete computer program based on this algorithm would be extremely long and time consuming, such a program could be very economical. Short, fast programs are available, but relative costs cannot be used to measure the worth of an essential result which may be obtainable in no other way.

1. Provide as "input" subroutines for evaluation of P(x) and P'(x) , and input computing parameters.

2. Determine a region, T , containing all roots.

3. Select appropriate starting values, if required, and apply an ordinary method to determine n approximate roots.

4. Determine an error bound for each approximate root.

5. Select the approximate root with the "best" error bound.

6. Iterate to completion the process for improving the error bound.

7. Repeat Steps 5 and 6 for all approximate roots which satisfy error bound criteria.

8. Select the best approximate root available.

9. If the "box" selected does not intersect any other box, count roots. If it does intersect another box, go to Step 12.

10. Conjugate roots if appropriate.

11. If all roots are located, go to Step 16. If all roots which satisfy error bound criteria are located but other roots remain, go to Step 15. Otherwise go to Step 8.

12. Determine the box which contains completely the original box and any others necessary to form a new box disjoint from all remaining boxes.

13. Count the roots in the new box. If there is only one, take the intersection of the original boxes and go to Step 10.

14. Subdivide the new box and count the roots in each partition. Continue until the roots are isolated or the process fails. Then go to Step 10.

15. If it is possible to use roots already determined to factor the original polynomial, return to Step 2 with the reduced polynomial.

16. Output those roots determined successfully together with complete error bound, and stop.

REFERENCES

REFERENCES

[1] P. S. Dwyer, Linear Computations, New York, 1951.

[2] R. E. Moore, Automatic Error Analysis in Digital Computation, LMSD-48421, Lockheed Missiles & Space Company, Sunnyvale, California, 1959.

[3] R. E. Moore, Interval Arithmetic and Automatic Error Analysis in Digital Computing, Technical Report number 25, Stanford University, 1962.

[4] R. E. Moore, J. A. Davison, H. R. Jaschke, and S. Shayer, DIFEQ Integration Routine – User's Manual, Technical Report LMSC 6-90-64-6, Lockheed Missiles & Space Company, Palo Alto, California, 1964.

[5] R. E. Moore, "The Automatic Analysis and Control of Error in Digital Computation Based on the Use of Interval Numbers." In Error in Digital Computation, Volume I, L. B. Rall (Editor), New York, 1965.

[6] R. E. Moore, "Automatic Local Coordinate Transformations to Reduce the Growth of Error Bounds in Interval Computation of Solutions of Ordinary Differential Equations." In Error in Digital Computation, Volume II, L. B. Rall (Editor), New York, 1965.

[7] P. S. Dwyer, Matrix Inversion with the Square Root Method, Technometrics, Volume 6, number 2 (1964).

[8] E. R. Hansen, Interval Arithmetic in Matrix Computations, Part I, Journal of Siam, Series B, Volume 2, number 2 (1965).

[9] R. E. Boche, "An Operational Interval Arithmetic," delivered at IEEE National Electronics Conference, Chicago, 1963.

[10] R. E. Boche, Specifications for an Interval Input Program, submitted (1965) to ACM.

[11] R. E. Boche, Some Observations on the Economics of Interval Arithmetic, Communications of the ACM, Volume 8, number 11 (1965).

[12] S. Shayer, Interval Arithmetic with Some Applications for Digital Computers, unpublished master's thesis, San Jose State College, 1965.

[13] K. Knopp, Elements of the Theory of Functions, New York, 1952.

[14] P. Henrici, "Elementary Numerical Analysis," lecture notes prepared for use at the Summer Institute for Numerical Analysis, University of California at Los Angeles, 1962.