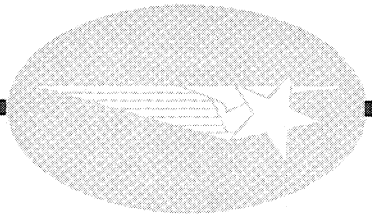
Eldon Hansen



Lockheed

MISSILES & SPACE COMPANY

A GROUP DIVISION OF LOCKHEED AIRCRAFT CORPORATION

SUNNYVALE, CALIFORNIA

INTERVAL ARITHMETIC WITH SOME APPLICATIONS FOR DIGITAL COMPUTERS

Sidney Shayer

5-13-65-12 July 1965

LOCKHEED MISSILES & SPACE COMPANY
Lockheed Palo Alto Research Laboratory
Palo Alto, California

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Purpose and Scope

The advent of high-speed digital computers have had a profound effect upon numerical analysis and on computational techniques. Problems that could not previously be done, because of the lengthy calculations involved, can now be completed in a relatively short period of time.

This tremendous advance in computing speed has brought with it problems relating to the accuracy of the results of lengthy machine computations. Not all of these problems are new, but many are now relatively more important than they were when calculations were done with paper and pencil or with the aid of a desk calculator. Decisions based on experience and insight into the particular problem being solved are no longer possible. All conditions should be planned for in advance when one does a problem on a digital computer.

The particular digital computer and programming system being used must be considered in a discussion of error analysis. The manner in which arithmetic is performed, numbers are represented, and in which round-off, truncation and significant errors are handled vary from computer to computer. Conversion from decimal to binary often forces the computation to be started with a value which only approximates the actual known value. The converting of irrational numbers to finite rational approximations adds to the problem of determining the accuracy of results.

Hence, it is exceedingly difficult to make a comprehensive study of the total error, even for a rather simple problem done on a digital computer. The net result is that rigorous error bounds are not usually known by, or available to, the user of a digital computer.

The purpose of this paper is to present a proposed solution to the problem of finding rigorous error bounds, by the use of interval arithmetic. Interval arithmetic is a method of computation in which the rational operations on real numbers are replaced by corresponding operations on closed intervals. That such a replacement is always possible will be shown in Chapter II, where a discussion is given of interval arithmetic as a system – definitions, operations, theorems, and elementary consequences.

In Chapter III, an algorithm is presented for the computation of certain functions by the Taylor series, and which has error bounds built into the algorithm. Finally, in Chapter IV, we employ closed intervals to solve the initial-value first-order differential equation

$$\frac{dy}{dx} = f(x, y) \tag{1.1}$$

where

$$y(x_o) = y_o \tag{1.2}$$

Historical Remarks

The concept of an interval arithmetic seems to have been first formulated by Paul S. Dwyer [1][1] and used by him for computation done on a desk calculator, but not on a digital computer. In 1951, Dwyer discussed, in a heuristic manner, certain methods for addition, subtraction, multiplication, division and the square root of his so-called "range numbers."

In 1954, Saul Gorn wrote an article [2] concerned with errors which derive from the following conditions:

---

[1]Numbers in brackets refer to references listed at the end of this paper.

(1) An initial inaccuracy is introduced by the necessary replacement of real numbers by finite decimal (or binary) approximations.
(2) Multiplications, divisions, and similar operations performed on two s-placed numbers in general yield s-placed results only if rounded off, with an accompanying "round-off error."
(3) Functions defined by infinite processes are computed by finitely truncated processes.

While Gorn does not mention interval arithmetic per se, the methods and procedures he describes would lead one naturally to construct an interval arithmetic. He realized that this was the case, for he wrote:

> ...it is possible to code an error-estimating routine applicable without change to any computational procedure whatever.

However, it was not until 1959 that formal proofs were presented, when R. E. Moore published a report [3] in which he presented a formal system, called interval arithmetic. George Collins (of the International Business Machines Corporation) published several interval arithmetic programs [4] in 1960, for the IBM 704 computer. Dr. Collins presented primarily an heuristic approach to the subject.

In 1962, Dr. Moore, working at Stanford University, produced another report [5] on interval arithmetic in which he states that an interval arithmetic computer program exists which could be used for the numerical solution of the initial-value problem for ordinary differential equations and would yield both rigorous and sharp error bounds.

Four papers published in 1964 in mathematical journals ([6], [7], [8], and [9]) indicate that several Russian mathematicians have also been working in the area of interval arithmetic for the numerical solution of differential equations.

CHAPTER II

INTERVAL ARITHMETIC

We shall now construct an algebraic system known as interval arithmetic. It will be seen that this system is an abelian semi-group and contains the field of real numbers.

Definitions and Operations

Definition 2.1: For any pair of real constants a, b, with $a \leq b$, the set of all real numbers x for which $a \leq x \leq b$ is called the closed interval $[a, b]$.

Since, corresponding to each pair of real constants a, b $(a \leq b)$ there exists a closed interval, the set of all closed intervals is necessarily infinite. Let $[a, b]$ denote a closed interval corresponding to a pair of constants a, b $(a \leq b)$, and let S denote the set of all closed intervals. The following definition may then be made.

Definition 2.2: Two closed intervals are said to be equal, i.e. $[a, b] = [c, d]$, if and only if $a = c$ and $b = d$.

We now proceed briefly to define [1] the arithmetical operations of two closed intervals $[a, b]$ and $[c, d]$ in S.

Definition 2.3: Interval addition:

$$[a, b] + [c, d] = [a + c, b + d] . \tag{2.1}$$

Definition 2.4: Interval multiplication:

$$[a, b][c, d] = [\min(ac, ad, bc, bd), \max(ac, ad, bc, bd)] . \tag{2.2}$$

Hereafter we shall refer to these operations simply as addition and multiplication when no confusion exists as to whether we are discussing interval arithmetic or the arithmetic of real numbers.

Fundamental Theorems

We will now present some simple theorems concerning interval arithmetic which will be needed in later chapters. It will first be demonstrated that interval arithmetic is closed, associative, and commutative under both operations of addition and multiplication. By virtue of our definition for a closed interval we must logically assume here the properties of real numbers.

THEOREM 2.1. (CLOSURE FOR ADDITION). For every pair of elements $[a,b]$ and $[c,d]$ in S there exists a unique element

$$[e,f] = [a,b] + [c,d] \tag{2.3}$$

in S.

PROOF: Let $[a,b]$ and $[c,d]$ be any two elements in S. Then we have

$$[a,b] + [c,d] = [a + c, b + d] \tag{2.4}$$

by virtue of Definition 2.3. Now since a, b, c, d are real numbers, so also are $a + c$ and $b + d$. Moreover, since $a \leq b$ and $c \leq d$ by Definition 2.1, so is $a + c \leq b + d$. This shows that $[a + c, b + d]$ is a closed interval and hence is an element of S. The uniqueness of $[a + c, b + d]$ follows from the uniqueness of the real numbers $a + c$, $b + d$. This completes the proof.

The proof of closure for multiplication is similar to the proof that addition is closed.

THEOREM 2.2. (CLOSURE FOR MULTIPLICATION). If $[a,b]$ and $[c,d]$ are any pair of elements in S, there exists a unique element

$$[e,f] = [a,b][c,d] \tag{2.5}$$

in S.

PROOF: Let $[a,b]$ and $[c,d]$ be any two elements in S. Then we have

$$[a,b][c,d] = [\min(ac,ad,bc,bd), \max(ac,ad,bc,bd)] \tag{2.6}$$

by virtue of Definition 2.4. Now since a,b,c,d are real numbers, so are the numbers ac,ad,bc,bd. Furthermore, $\min(ac,ad,bc,bd) \leq \max(ac,ad,bc,bd)$, which shows that

$$[\min(ac,ad,bc,bd), \max(ac,ad,bc,bd)]$$

is a closed interval and hence is an element of S. The uniqueness of the interval $[\min(ac,ad,bc,bd), \max(ac,ad,bc,bd)]$ follows from the uniqueness of the real numbers ac,ad,bc,bd. The proof of the theorem is thus complete.

It will now be shown that addition and multiplication are associative.

THEOREM 2.3. (ASSOCIATIVITY FOR ADDITION). For all elements $[a,b]$, $[c,d]$, $[e,f]$ in S,

$$[a,b] + ([c,d] + [e,f]) = ([a,b] + [c,d]) + [e,f] . \tag{2.7}$$

PROOF: Let $[a,b]$, $[c,d]$, $[e,f]$ be any three elements in S. Then we have, by Definition 2.3,

$$\begin{aligned} [a,b] + ([c,d] + [e,f]) &= [a,b] + [c+e, d+f] \\ &= [a+c+e, b+d+f] . \end{aligned} \tag{2.8}$$

Also, by Definition 2.3, it is seen that

$$\begin{aligned} ([a,b] + [c,d]) + [e,f] &= [a+c, b+d] + [e,f] \\ &= [a+c+e, b+d+f] . \end{aligned} \tag{2.9}$$

Thus, summing each side of (2.7) yields the same closed interval, and the proof is complete.

THEOREM 2.4. (ASSOCIATIVITY FOR MULTIPLICATION). For all elements $[a,b]$, $[c,d]$, $[e,f]$ in S ,

$$[a,b]\,([c,d][e,f]) = ([a,b][c,d])\,[e,f] \quad . \tag{2.10}$$

PROOF: Let $[a,b]$, $[c,d]$, $[e,f]$ be any three elements in S . Then by Definition 2.4, we have for the left side of (2.10),

$$\begin{aligned}[a,b]\,([c,d][e,f]) &= [a,b][\min(ce,cf,de,df), \max(ce,cf,de,df,)] \\ &= [\min(ace,acf,ade,adf,bce,bcf,bde,bdf), \\ &\quad \max(ace,acf,ade,adf,bce,bcf,bde,bdf)] \quad .\end{aligned} \tag{2.11}$$

The right side of equality (2.10) yields, by Definition 2.4,

$$\begin{aligned}([a,b][c,d])\,[e,f] &= [\min(ac,ad,bc,bd), \max(ac,ad,bc,bd)][e,f] \\ &= [\min(ace,acf,ade,adf,bce,bcf,bde,bdf), \\ &\quad \max(ace,acf,ade,adf,bce,bcf,bde,bdf)] \quad .\end{aligned} \tag{2.12}$$

Hence, multiplying each side of (2.10) gives the same closed interval, and the theorem is proved.

The following two theorems show that addition and multiplication are commutative.

THEOREM 2.5. (COMMUTATIVITY FOR ADDITION). For every pair of elements $[a,b]$ and $[c,d]$ in S the relation

$$[a,b] + [c,d] = [c,d] + [a,b] \tag{2.13}$$

holds.

PROOF: Let $[a,b]$ and $[c,d]$ be any two elements in S . Then we have, by Definition 2.3,

$$[a,b] + [c,d] = [a+c, b+d] \tag{2.14}$$

and

(2.15) $$[c,d] + [a,b] = [c + a, d + b] .$$

Since a,b,c,d are real numbers, so also are $a + c$, $b + d$, $c + a$, $d + b$, and commutativity holds for the addition of real numbers. Hence $a + c = c + a$ and $b + d = d + b$. Therefore,

(2.16) $$[a + c, b + d] = [c + a, d + b] ,$$

which implies that (2.13) is an identity and thus completes our proof.

THEOREM 2.6. (COMMUTATIVITY FOR MULTIPLICATION). If $[a,b]$ and $[c,d]$ are any pair of elements in S , then

(2.17) $$[a,b][c,d] = [c,d][a,b] .$$

PROOF: Let $[a,b]$ and $[c,d]$ be any two elements in S . Then by Definition 2.4, we have

(2.18) $$[a,b][c,d] = [\min(ac, ad, bc, bd) , \max(ac, ad, bc, bd)]$$

and

(2.19) $$[c,d][a,b] = [\min(ca, cb, da, db) , \max(ca, cb, da, db)] .$$

Since a,b,c,d are real numbers, so are their products, and the real numbers are commutative under multiplication. Hence $ac = ca$, $ad = da$, $bc = cb$, $bd = db$, and it is therefore seen that the right side of equation (2.18) equals the right side of equation (2.19). This implies that (2.17) is an identity, and our proof is complete.

We now show that the closed intervals $[0,0]$ and $[1,1]$ are both left and right identities for addition and multiplication, respectively.

THEOREM 2.7. (IDENTITY FOR ADDITION). The closed interval $[0,0]$ is both a left and right identity for addition, i.e.,

$$[0,0] + [a,b] = [a,b] + [0,0] = [a,b] . \tag{2.20}$$

PROOF: Let $[a,b]$ be any element in S. By THEOREM 2.5, viz. commutativity for addition, we have

$$[0,0] + [a,b] = [a,b] + [0,0] . \tag{2.21}$$

Now, one sees by Definition 2.3 that

$$[0,0] + [a,b] = [0 + a , 0 + b] , \tag{2.22}$$

and that

$$[a,b] + [0,0] = [a + 0 , b + 0] . \tag{2.23}$$

Since $0, a, b$ are real numbers, their sums are also real. In particular, $0 + a = a$, $a + 0 = a$, $0 + b = b$, and $b + 0 = b$, so that

$$[0,0] + [a,b] = [0 + a , 0 + b] = [a,b] \tag{2.24}$$

and

$$[a,b] + [0,0] = [a + 0 , b + 0] = [a,b] . \tag{2.25}$$

Therefore, the closed interval $[0,0]$ is both a left and right identity for addition.

THEOREM 2.8. (IDENTITY FOR MULTIPLICATION). The closed interval $[1,1]$ is both a left and right identity for multiplication, i.e.,

$$[1,1][a,b] = [a,b][1,1] = [a,b] . \tag{2.26}$$

PROOF: Let $[a,b]$ be any element in S. By THEOREM 2.6 for the commutativity for multiplication, we have

(2.27) $$[1,1][a,b] = [a,b][1,1] .$$

Now, by Definition 2.4,

(2.28) $$[1,1][a,b] = [\min(1 \cdot a, 1 \cdot b), \max(1 \cdot a, 1 \cdot b)] ,$$

and

(2.29) $$[a,b][1,1] = [\min(a \cdot 1, b \cdot 1), \max(a \cdot 1, b \cdot 1)] .$$

Since $1, a, b$ are real numbers, their products are real. In particular, $1 \cdot a = a$, $a \cdot 1 = a$, $1 \cdot b = b$, $b \cdot 1 = b$. Remembering that $a \leq b$, we have

(2.30) $$[\min(1 \cdot a, 1 \cdot b), \max(1 \cdot a, 1 \cdot b)] = [a,b]$$

and

(2.31) $$[\min(a \cdot 1, b \cdot 1), \max(a \cdot 1, b \cdot 1)] = [a,b] .$$

Hence, it follows that the closed interval $[1,1]$ is both a left and right identity for multiplication.

Elementary Consequences

Some elementary consequences, based on the definitions and theorems will now be presented.

The reflexive, symmetric, and transitive laws hold for all $[a,b]$, $[c,d]$, $[e,f]$ in S and are immediate consequences of real numbers. We thus have the following.

Reflexive law:

(2.32) $$[a,b] = [a,b] .$$

Symmetric law:

(2.33) If $[a,b] = [c,d]$, then $[c,d] = [a,b]$.

Transitive law:

(2.34) If $[a,b] = [c,d]$, and $[c,d] = [e,f]$,

then $[a,b] = [e,f]$.

It is of interest to note that an isomorphism[1] exists between the field of real numbers and closed intervals of the form $[a,a]$.

Definition 2.5: The set S^* is the set composed of all closed intervals of the form $[a,a]$, where a is a real number.

Closed intervals of the form $[a,a]$ are intervals of zero width. Clearly S^* is a subset of S , and the set S^* is necessarily infinite.

THEOREM 2.9. (S^* ISOMORPHIC TO THE REAL NUMBERS). The set of S^* is isomorphic to the field of real numbers.

PROOF: We first establish a one-one correspondence between the field of real numbers and the elements of the set S^* . We do this in the following manner: with 0 we associate $[0,0]$, and vice versa, while with 1 we associate $[1,1]$, and vice versa; in general, we associate n with $[n,n]$, and vice versa.

We now show that sums and products are preserved. Let a and b be any two real numbers. Then by the one-one correspondence established above,

---

[1] An isomorphism between two sets S and S' is a one-one correspondence $a \longleftrightarrow a'$ of the elements of S with the elements of S' , which satisfies for all a and b the conditions

$$(a + b)' = a' + b' \quad , \quad (ab)' = a'b' \ .$$

the number a corresponds to the element $[a,a]$ in S^*, and b corresponds to the element $[b,b]$ in S^*. Hence, we must show that $a+b$ corresponds to the sum of the closed intervals $[a,a]+[b,b]$, and that ab corresponds to the product of the closed intervals $[a,a][b,b]$. From Definition 2.3 we have

$$[a,a] + [b,b] = [a+b, a+b] . \tag{2.35}$$

and by Definition 2.4 it follows that

$$[a,a][b,b] = [ab, ab] . \tag{2.36}$$

By the one-one correspondence established above

$$a + b \longleftrightarrow [a+b, a+b] , \tag{2.37}$$

and

$$ab \longleftrightarrow [ab, ab] \tag{2.38}$$

Thus, sums and products are preserved and the isomorphism is proved.

The field of real numbers can be considered as being embedded within the set S of closed intervals. By the correspondence

$$c \longleftrightarrow [c,c] \tag{2.39}$$

any real number may be changed to a closed interval. Thus we can define addition and multiplication of real numbers and closed intervals.

Definition 2.6: Addition of a real number c and a closed interval $[a,b]$ in S is defined by the relation

$$c + [a,b] = [c+a, c+b] . \tag{2.40}$$

Definition 2.7: Multiplication of a real number, c, and a closed interval $[a,b]$ in S is defined by

$$c[a,b] = [\min(ca,cb), \max(ca,cb)] . \tag{2.41}$$

In general, neither additive nor multiplicative inverses exists in the set of closed intervals S .

THEOREM 2.10. (ADDITIVE INVERSES). Additive inverses do not exist in the set of closed intervals S , except for the subset S^* .

PROOF: The set S^* is isomorphic to the real numbers by THEOREM 2.9, and additive inverses exist for all elements of the set of real numbers. Additive inverses will therefore exist for all elements of S^* .

Let $[a,b]$ be any closed interval in S , where $a \neq b$ (i.e., $a < b$). Assume that an additive inverse exists for $[a,b]$ in S , say $[c,d]$. Then by Definition 2.3

$$[a,b] + [c,d] = [a+c, b+d] = [0,0] , \tag{2.42}$$

and therefore,

$$a + c = 0 , \quad \text{and} \quad b + d = 0 , \tag{2.43}$$

Thus,

$$a = -c , \quad \text{and} \quad b = -d . \tag{2.44}$$

Since $a < b$ by hypothesis, we have

$$a = -c < b = -d \quad \text{and} \quad -c < -d , \tag{2.45}$$

This implies that

$$d < c ; \tag{2.46}$$

but by Definition 2.1, $[c,d]$ requires that

$$c \leq d . \tag{2.47}$$

Thus, assuming that $[a,b]$ has an additive inverse leads us to a contradiction, and the theorem is proved.

THEOREM 2.11. (MULTIPLICATIVE INVERSES). Multiplicative inverses do not exist in the set of closed intervals S, except for a subset S^* which excludes $[0,0]$.

PROOF: The set S^* is isomorphic to the real numbers by THEOREM 2.9, and multiplicative inverses exist for all elements of the set of real numbers, except zero. Therefore, multiplicative inverses exist for all elements of S^*, except $[0,0]$.

Let $[a,b]$ be any closed interval in S, where $a \neq b$, (i.e., $a < b$). Suppose a multiplicative inverse exists for $[a,b]$ in S, say $[c,d]$. Then by Definition 2.4, we have

$$[a,b][c,d] = [\min(ac, ad, bc, bd), \max(ac, ad, bc, bd)] = [1,1], \tag{2.48}$$

implying that

$$1 \leq ac \leq 1, \quad 1 \leq ad \leq 1, \quad 1 \leq bc \leq 1, \quad 1 \leq bd \leq 1. \tag{2.49}$$

But if

$$ac = bc \quad \text{or} \quad ad = bd \tag{2.50}$$

then

$$a = b. \tag{2.51}$$

This, however, contradicts our assumption that $a < b$, and the proof is complete.

It is interesting to note that even though additive inverses do not exist for all elements of S, we have the following theorem:

THEOREM 2.12. (CANCELLATION LAW FOR ADDITION). For $[a,b]$, $[c,d]$, $[e,f]$ in S, if

$$[a,b] + [c,d] = [a,b] + [e,f], \tag{2.52}$$

then

(2.53) $$[c,d] = [e,f] .$$

PROOF: Let $[a,b]$, $[c,d]$, $[e,f]$ be elements of S . By assumption,

(2.54) $$[a,b] + [c,d] = [a,b] + [e,f] .$$

Summing each side of equation (2.54) yields

(2.55) $$[a + c , b + d] = [a + e , b + f] .$$

From Definition 2.2, we have

(2.56) $$a + c = a + e \quad , \quad \text{and} \quad b + d = b + f .$$

Recalling that a, b, c, d, e, f are real numbers, we note that equation (2.56) implies that

(2.57) $$c = e \quad , \quad \text{and} \quad d = f .$$

Hence, by Definition 2.2, it follows that

(2.58) $$[c,d] = [e,f] .$$

THEOREM 2.13. The cancellation law for multiplication does not hold in interval arithmetic. That is, for the intervals $[a,b]$, $[c,d]$, $[e,f]$ in S , if

(2.59) $$[a,b][c,d] = [a,b][e,f] ,$$

we cannot conclude that

(2.60) $$[c,d] = [e,f] .$$

PROOF: (Let us assume that the cancellation law does hold in interval arithmetic. We shall then show by a counter example that this assumption is incorrect). For the closed intervals $[0,2]$, $[0,1]$, $[1,1]$ in S , it is seen

by Definition 2.4 that

$$[0,2][0,1] = [0,2][1,1] \tag{2.61}$$

$$[0,2] = [0,2] .$$

Now

$$[0,1] \neq [1,1] \tag{2.62}$$

and thus our assumption that the cancellation law holds is incorrect, and our theorem is proved.

Another point of interest concerns the distributive law.

THEOREM 2.14. The distributive law does not hold in interval arithmetic.

PROOF: Let $[a,b]$, $[c,c]$, $[-c,-c]$ be in S . Using Definitions 2.3 and 2.4, and first adding and then multiplying, yields

$$[a,b]([c,c] + [-c,-c]) = [a,b][0,0] = [0,0] . \tag{2.63}$$

But if we first multiply and then add, we get

$$\begin{aligned} [a,b]([c,c] + [-c,-c]) &= [a,b][c,c] + [a,b][-c,-c] \\ &= [\min(ac,bc), \max(ac,bc)] \\ &\quad + [\min(-ac,-bc), \max(-ac,-bc)] \\ &\neq 0 \end{aligned} \tag{2.64}$$

unless $a = b = 0$, or $c = -c = 0$, or both. Therefore, in the general case, the distributive law does not hold in interval arithmetic.

One special case in which the distributive law does hold, however, is noteworthy:

THEOREM 2.15. If $[a,a]$ is in S^* and $[b,c]$ and $[d,e]$ are in S , then

$$[a,a]([b,c] + [d,e]) = [a,a][b,c] + [a,a][d,e] . \tag{2.65}$$

PROOF: Let $[a,a]$ be any element in S^*, and let $[b,c]$ and $[d,e]$ be any two elements in S. There are three cases, depending on whether $a < 0$, $a = 0$, $a > 0$.

Case 1. $a < 0$:

$$[a,a]([b,c] + [d,e]) = [a,a][b,c] + [a,a][d,e] \tag{2.66}$$

$$[a,a][b+d, c+e] = [a,a][b,c] + [a,a][d,e]$$

$$[a(c+e), a(b+d)] = [ac, ab] + [ae, ad]$$

and finally

$$[ac + ae, ab + ad] = [ac + ae, ab + ad].$$

Case 2. $a = 0$:

$$[0,0]([b,c] + [d,e]) = [0,0][b,c] + [0,0][d,e] \tag{2.67}$$

$$[0,0][b+d, c+e] = [0,0] + [0,0]$$

which yields

$$[0,0] = [0,0].$$

Case 3. $a > 0$:

$$[a,a]([b,c] + [d,e]) = [a,a][b,c] + [a,a][d,e] \tag{2.68}$$

$$[a,a][b+d, c+e] = [a,a][b,c] + [a,a][d,e]$$

$$[a(b+d), a(c+e)] = [ab, ac] + [ad, ae]$$

and so

$$[ab + ad, ac + ae] = [ab + ad, ac + ae].$$

In each case the equality of (2.65) holds, and the proof is complete.

Hence, the distributive law holds when the common factor is an element of S^*. Due to the isomorphism between the real numbers and the set S^* (THEOREM 2.9), and to Definitions 2.6 and 2.7, which give us the ability to

both add and multiply closed intervals by real numbers, it is clear from THEOREM 2.15 that

$$n([a,b] + [c,d]) = n[a,b] + n[c,d] \tag{2.69}$$

where n is a real number and $[a,b]$, $[c,d]$ are in S.

The elements of S can be partially ordered [2] by set inclusion.

Definition 2.8: A partial ordering by set inclusion for any two elements $[a,b]$ and $[c,d]$ in S, exists if and only if

$$c \leq a \leq b \leq d . \tag{2.70}$$

This ordering is expressed symbolically by

$$[a,b] \subseteq [c,d] . \tag{2.71}$$

It follows from Definition 2.8 that

$$[a,b] \subseteq [c,d] \quad \text{and} \quad [c,d] \subseteq [a,b] \tag{2.72}$$

if and only if

$$a = c \quad \text{and} \quad b = d . \tag{2.73}$$

It is clear that for the closed intervals $[a,b]$, $[c,d]$, $[e,f]$, $[g,h]$ in S, if

$$[a,b] \subseteq [e,f] \quad \text{and} \quad [c,d] \subseteq [g,h] , \tag{2.74}$$

then

$$[a,b] + [c,d] \subseteq [e,f] + [g,h] \tag{2.75}$$

and

$$[a,b][c,d] \subseteq [e,f][g,h] . \tag{2.76}$$

The inclusion relationships follow from the definitions of addition and multiplication, respectively, viz., Definitions 2.3 and 2.4.

While in the general case distributivity does not hold for closed intervals, a weaker law, which Moore [3] calls "subdistributivity," does hold.

Definition 2.9: Subdistributivity: for the closed intervals $[a,b]$, $[c,d]$, $[e,f]$ in S:

$$[a,b]([c,d]+[e,f]) \subseteq [a,b][c,d]+[a,b][e,f] . \tag{2.77}$$

For example,

$$[1,2]([3,3]+[-3,-3]) \subset [1,2][3,3]+[1,2][-3,-3]$$

$$[1,2][0,0] \subset [3,6]+[-6,-3]$$

$$[0,0] \subset [-3,3] .$$

The usual power notation will be used.

Definition 2.10: For any closed interval $[a,b]$ in S, and any integer $n \geq 0$,

$$[a,b]^n = [a,b][a,b] \ldots [a,b] , \tag{2.78}$$

where the closed interval $[a,b]$ is used as a factor n times, and for $n = 0$ we mean

$$[a,b]^0 = [1,1] . \tag{2.79}$$

The concepts of union and intersection have meaning with closed intervals as in the following:

Definition 2.11: The union of two closed intervals, $[a,b]$ and $[c,d]$ in S, is defined by the relation

$$[a,b] \cup [c,d] = \{x | x \in [a,b] \text{ or } x \in [c,d]\} . \tag{2.80}$$

Definition 2.12: The intersection of two closed intervals, $[a,b]$ and $[c,d]$ in S, is defined by

$$[a,b] \cap [c,d] = \{x | x \in [a,b] \text{ and } x \in [c,d]\} . \tag{2.81}$$

Besides the partial ordering by set inclusion (Definition 2.8), the set S can be partially ordered by the inequality, $<$, less than [4].

Definition 2.13: A partial ordering by the inequality, $<$, less than, for the closed intervals $[a,b]$ and $[c,d]$ in S is defined by

$$[a,b] < [c,d] \tag{2.82}$$

where for every $x \in [a,b]$ and for every $y \in [c,d]$,

$$x < y \ . \tag{2.83}$$

The ordering of closed intervals by Definition 2.13 is a transitive relationship. Since for $x \in [a,b]$, $y \in [c,d]$, and $z \in [e,f]$, if

$$[a,b] < [c,d] \quad \text{and} \quad [c,d] < [e,f] \tag{2.84}$$

then by Definition 2.13

$$x < y \quad \text{and} \quad y < z \ . \tag{2.85}$$

But x,y,z are real and the real numbers are ordered. Further, the ordering of the real numbers is a transitive relationship. Thus (2.84) implies that

$$[a,b] < [e,f] \ . \tag{2.86}$$

Definition 2.14: The square root of the interval $[a,b]$ in S, with $a \geq 0$, is defined by

$$\sqrt{[a,b]} = [\sqrt{a}, \sqrt{b}] \ . \tag{2.87}$$

For $\sqrt{[a,b]}$, we wish to obtain an interval which contains the square root of each x in the closed interval

$$a \leq x \leq b \ , \quad a \geq 0 \ . \tag{2.88}$$

Now the least square root of (2.87) will be $\sqrt{a}$ and the greatest square root of the interval will be $\sqrt{b}$. Thus the interval $[\sqrt{a}, \sqrt{b}]$ will contain all values of the square root of x contained in $[a,b]$.

An example should clarify this concept. Let us obtain the square root of the closed interval $[1,2]$. That is, we want the square root of x , say to three decimal places, where $1 \leq x \leq 2$. Now, by (2.87),

$$\sqrt{[1,2]} = [\sqrt{1}, \sqrt{2}] = [1, 1.415] \ .$$

Clearly, the closed interval $[1, 1.415]$ contains the square root of all values of x in the given closed interval, $[1,2]$.

We shall now define subtraction and division for two closed intervals $[a,b]$ and $[c,d]$ in S .

Definition 2.15: Interval subtraction:

$$[a,b] - [c,d] = [a,b] + [-d, -c] \ . \tag{2.89}$$

Definition 2.16: Interval Division:

$$[a,b] \div [c,d] = [a,b][1/d, 1/c] \ , \tag{2.90}$$

provided $0 \notin [c,d]$.

We should observe that the interval fraction

$$\frac{[a,b]}{[a,b]} \neq [1,1] \quad , \quad 0 \notin [a,b] \tag{2.91}$$

unless $a = b$. For, by Definition 2.16

$$\frac{[a,b]}{[a,b]} = [a,b]\left[\frac{1}{b}, \frac{1}{a}\right] = \left[\min\left(\frac{a}{b}, 1, \frac{b}{a}\right), \max\left(\frac{a}{b}, 1, \frac{b}{a}\right)\right] \ . \tag{2.92}$$

Only in the case when $a = b$ do we have

$$\frac{[a,b]}{[a,b]} = [1,1] \ . \tag{2.93}$$

If $0 < a < b$, then

$$\frac{[a,b]}{[a,b]} = \left[\frac{a}{b}, \frac{b}{a}\right] \quad . \tag{2.94}$$

If $a < b < 0$, then

$$\frac{[a,b]}{[a,b]} = \left[\frac{b}{a}, \frac{a}{b}\right] \quad . \tag{2.95}$$

If $a < 0 < b$, we have an interval that contains zero and we cannot divide.

Definition 2.17: A rational interval expression [5]

$$F([x_1, x_2], [x_3, x_4], \ldots, [x_{n-1}, x_n]) \tag{2.96}$$

is a finite combination of closed interval variables,

$$[x_1, x_2], [x_3, x_4], \ldots, [x_{n-1}, x_n]$$

and a finite set of constant closed intervals of the form $[a,b]$ in an expression with interval arithmetic operations.

A rational interval form is usually not representable as a quotient of two polynomials. That is, we cannot say

$$[x_1, x_2] + \frac{[1,1]}{[x_1, x_2]} = \frac{[x_1, x_2]^2 + [1,1]}{[x_1, x_2]} , \tag{2.97}$$

unless $x_1 = x_2$. This is due to (2.91).

Since interval arithmetic operations are monotonic inclusive, if

$$[x'_1, x'_2] \subset [x_1, x_2], [x'_2, x'_3] \subset [x_3, x_4], \ldots, \tag{2.98}$$

$$[x'_{n-1}, x'_n] \subset [x_{n-1}, x_n]$$

and if

(2.99) $$F([x_1, x_2], [x_3, x_4], \ldots, [x_{n-1}, x_n])$$

is a rational interval expression, then

(2.100) $$F([x'_1, x'_2], [x'_3, x'_4], \ldots, [x'_{n-1}, x'_n]) \subset F([x_1, x_2], [x_3, x_4], \ldots, [x_{n-1}, x_n]) .$$

The following two theorems are of value when doing numerical calculations with closed intervals.

THEOREM 2.16. For the closed intervals $[a,b]$, $[c,d]$ in S, if

(2.101) $$[a,b] \subset [c,d]$$

then there exists a closed interval $[e,f]$ in S, such that

(2.102) $$[c,d] = [a,b] + [e,f] ,$$

and $0 \in [e,f]$.

PROOF: By hypothesis,

(2.103) $$[a,b] \subset [c,d] ,$$

So that from Definition 2.8,

(2.104) $$c \leq a \leq b \leq d .$$

Since a,b,c,d are real numbers, for $c \leq a$ and $b \leq d$, some real numbers $e \leq 0$, and $f \geq 0$ exist such that

(2.105) $$c = a + e \quad \text{and} \quad d = b + f$$

Thus,

(2.106) $$[c,d] = [a,b] + [e,f] ,$$

for, by Definition 2.3

$$[c,d] = [a+e, b+f] . \tag{2.107}$$

THEOREM 2.17. For the closed intervals $[a,b]$ and $[c,d]$ in S, if

$$[a,b] \subset [c,d] \tag{2.108}$$

and $[e,f]$ is any element in S which does not contain zero in its interior (zero can be an end point), then there exists a closed interval $[g,h]$ in S, such that

$$[c,d] = [a,b] + [e,f][g,h] . \tag{2.109}$$

PROOF: By hypothesis,

$$[a,b] \subset [c,d] , \tag{2.110}$$

and by THEOREM 2.16, we have

$$[c,d] = [a,b] + [x_1, x_2] , \ 0 \in [x_1, x_2]. \tag{2.111}$$

Let $[e,f]$ be any closed interval in S, which does not contain zero in its interior. Then we must show that a closed interval $[g,h]$ in S exists such that

$$[x_1, x_2] = [e,f][g,h] . \tag{2.112}$$

If $[g,h]$ exists, then by Definition 2.4

$$[e,f][g,h] = [\min(eg, eh, fg, fh) , \max(eg, eh, fg, fh)] . \tag{2.113}$$

Since by hypothesis $[e,f]$ does not contain zero in its interior, either

$$e \leq f \leq 0 \quad \text{or} \quad 0 \leq e \leq f \tag{2.114}$$

Since there are no restrictions on $[g,h]$, one of the following three conditions must hold for $[g,h]$:

$$g \leq h \leq 0 \quad , \quad g \leq 0 \leq h \quad , \quad 0 \leq g \leq h \; . \tag{2.115}$$

Thus, there are six possible cases for equation (2.113). We will show that in each case, $[g,h]$ exists, i.e., that equation (2.112) is satisfied.

Case 1. $e \leq f \leq 0$, $g \leq h \leq 0$:

$$[x_1, x_2] = [e,f][g,h] = [fh, eg] \; . \tag{2.116}$$

Case 2. $e \leq f \leq 0$, $g \leq 0 \leq h$:

$$[x_1, x_2] = [e,f][g,h] = [eh, eg] \; . \tag{2.117}$$

Case 3. $e \leq f \leq 0$, $0 \leq g \leq h$:

$$[x_1, x_2] = [e,f][g,h] = [eh, fg] \; . \tag{2.118}$$

Case 4. $0 \leq e \leq f$, $g \leq h \leq 0$:

$$[x_1, x_2] = [e,f][g,h] = [fg, eh] \; . \tag{2.119}$$

Case 5. $0 \leq e \leq f$, $g \leq 0 \leq h$:

$$[x_1, x_2] = [e,f][g,h] = [fg, fh] \; . \tag{2.120}$$

Case 6. $0 \leq e \leq f$, $0 \leq g \leq h$:

$$[x_1, x_2] = [e,f][g,h] = [eg, fh] \; . \tag{2.121}$$

Hence, in each case, a closed interval $[g,h]$ in S exists and the theorem is proved.

Interval Arithmetic As A Semi-Group

In this section we will prove that interval arithmetic is an abelian semi-group.

Definition 2.18: A set of elements which is closed under a binary operation and for which the associative law holds is called a semi-group [6]. If the operation is commutative, we have an abelian semi-group.

THEOREM 2.18. The set of closed intervals S forms an abelian semi-group under addition.

PROOF: By Definition 2.3, interval addition is a binary operation on the elements of S. THEOREMS 2.1, 2.3, and 2.5 give us closure, associativity, and commutativity, respectively, for the operation of addition in S. Therefore, the set S is an abelian semi-group under addition.

It should be noted that an abelian semi-group under addition is also known as a semi-module [7].

THEOREM 2.19. The set of closed intervals S, forms an abelian semi-group under multiplication.

PROOF: By Definition 2.4 interval multiplication is a binary operation on the elements of S. THEOREMS 2.2, 2.4, and 2.6 give us closure, associativity, and commutativity, respectively, for the operation of multiplication in S. The set S is therefore an abelian semi-group under multiplication.

The set S of closed intervals does not form an integral domain since the distributive and cancellation laws do not hold (THEOREMS 2.14 and 2.13, respectively) and we do not have an additive inverse (THEOREM 2.10).

Obviously S is not a field, for a field requires an integral domain with a multiplicative inverse, and by THEOREM 2.11, we do not have multiplicative inverses.

S is not a group under either addition or multiplication, for while we have identity elements for each (THEOREMS 2.7 and 2.8, respectively) we do

not have additive or multiplicative inverses (THEOREMS 2.10 and 2.11, respectively). Not being a group, S cannot be a ring.

In closing, it should be noted that we have by no means exhausted the study of interval arithmetic as an algebraic system, but have tried to present only the basic definitions and theorems and some useful elementary consequences. We now go on to some applications where interval arithmetic is useful in securing error bounds on numerical computations.

CHAPTER III

APPLICATION TO TAYLOR'S SERIES

Introduction

When a function $f(x)$ is approximated by a power series

$$f(x) \cong \sum_{i=o}^{n} a_i x^i \,, \tag{3.1}$$

the error ϵ is given by

$$f(x) - \sum_{i=o}^{n} a_i x^i = \epsilon \,. \tag{3.2}$$

If the error ϵ can be expressed analytically so that a reasonable upper bound may be found, and if the error term can be restricted to a region where it is monotonically increasing or decreasing, then by using interval arithmetic we can calculate values for the given function, and for each solution have a rigorous error bound. Moreover, the error bound would then be obtained simultaneously with the solution – not as a separate and time-consuming calculation.

As a practical matter, the techniques to be developed in this chapter are most useful if one is planning on having the calculations done on a digital computer. We shall discuss Taylor's series (since it is often used on digital computers for approximating functions), and shall place our emphasis on the remainder term of the series. It will be seen that the same general approach is applicable to many other series approximations.

Taylor's Theorem

Taylor's well-known theorem (sometimes called Taylor's formula, or Taylor's series with remainder) may be stated as follows [1]:

THEOREM 3.1. (TAYLOR'S THEOREM). Let a function $f(x)$ and its first n derivatives $(n \geq 0)$ be continuous in a closed interval containing $x = a$, and let x be any point in this interval. Then

$$f(x) = f(a) + (x - a)\, f'(a) + \frac{(x - a)^2}{2!} f''(a) + \ldots + \frac{(x - a)^{n-1}}{(n - 1)!} f^{(n-1)}(a) + R_n , \tag{3.3}$$

where the remainder R_n is given by

$$R_n = \frac{1}{(n - 1)!} \int_a^x (x - t)^{n-1} f^{(n)}(t)\, dt . \tag{3.4}$$

PROOF:[1] (By mathematical induction) By hypothesis, the given function $f(x)$ satisfies the Fundamental Theorem of the Integral Calculus, and hence

$$f(x) = f(a) + \int_a^x f'(t)\, dt . \tag{3.5}$$

This is formula (3.3) for $n = 1$. Assume now that (3.3) is true for n, that is, that

$$f(x) = f(a) + (x - a)\, f'(a) + \frac{(x - a)^2}{2!} f''(a) + \ldots + \frac{(x - a)^{n-1}}{(n - 1)!} f^{(n-1)}(a) + \frac{1}{(n - 1)!} \int_a^x (x - t)^{n-1} f^{(n)}(t)\, dt . \tag{3.6}$$

---

[1] In order to make the paper more self-contained, we give the proof here, although the theorem is proved in many standard texts.

Integrating the last term by parts, we obtain

$$f(x) = f(a) + (x - a)\, f'(a) + \frac{(x - a)^2}{2!} f''(a) + \ldots + \frac{(x - a)^{n-1}}{(n - 1)!} f^{(n-1)}(a) + \frac{(x - a)^n}{n!} f^{(n)}(a) + \frac{1}{n!} \int_a^x (x - t)^n f^{(n+1)}(t)\, dt \,, \tag{3.7}$$

which is the same formula as (3.3) with n replaced by $n + 1$. Therefore, by mathematical induction, formula (3.3) is true for all $n \geq 1$, and the proof is complete.

Of particular interest for our purposes is the remainder term

$$R_n = \frac{1}{(n - 1)!} \int_a^x f^{(n)}(t)(x - t)^{n-1}\, dt \,. \tag{3.8}$$

However, while this form of the remainder has the advantage of being explicit, it is usually rather difficult to estimate in a numerical problem. We will therefore develop Lagrange's form of the remainder

$$R_n = \frac{(x - a)^n}{n!} f^{(n)}(\xi) \,, \quad a < \xi < x \tag{3.9}$$

which has the advantage of simplicity and, as will be seen, is more amenable to interval arithmetic.

In order to obtain Lagrange's form of the remainder from (3.8), we make use of the Second Law of the Mean [2], which states that if $f(x)$ is continuous for $a \leq x \leq b$ and if $g(x)$ does not change sign for $a < x < b$, then the relation

$$(3.10) \qquad \int_a^b f(x)\, g(x)\, dx = f(\xi) \int_a^b g(x)\, dx$$

holds for at least one ξ such that $a < \xi < b$.

Now, the quantity $(x - t)^{n-1}$ in (3.8) does not change sign as t varies from a to x. Thus, the Second Law of the Mean allows us to rewrite (3.8) in the form

$$(3.11) \qquad R_n = \frac{f^{(n)}(\xi)}{(n-1)!} \int_a^x (x - t)^{n-1}\, dt \ , \quad a < \xi < x \ ,$$

which is easily integrated to yield

$$(3.12) \qquad R_n = \frac{(x - a)^n}{n!} f^{(n)}(\xi) \ , \quad a < \xi < x \ .$$

If we set $a = 0$ in (3.3), we get the well-known special case, called Maclaurin's series with remainder. That is,

$$(3.13) \qquad f(x) = f(o) + xf'(o) + \frac{x^2}{2!} f''(o) + \ldots + \frac{x^{n-1}}{(n-1)!} f^{(n-1)}(o) + R_n \ ,$$

with the remainder term (3.9) becoming

$$(3.14) \qquad R_n = \frac{x^n}{n!} f^{(n)}(\xi) \ , \quad o < \xi < x \ .$$

The Remainder Term as an Interval

The Lagrange form of the remainder (3.9) of Taylor's series tell us only that ξ is known to lie somewhere between a and x. Our objective is to obtain a closed interval which will contain the exact value of the remainder R_n.

For a function whose n^{th} derivative is either monotonically increasing or decreasing in the interval a to x , we can construct a closed interval such that the function can be approximated by Taylor's series with rigorous error bounds.

We construct the desired closed interval in the following manner. Noting that the remainder,

$$R_n = \frac{(x-a)^n}{n!} f^{(n)}(\xi) \quad , \quad a < \xi < x \quad , \tag{3.15}$$

states that the exact value is dependent upon ξ , and that ξ lies between a and x , we will build a closed interval which will contain all possible values of R_n , for $a < \xi < x$. If ξ takes on every possible value between a and x , then R_n is bounded. In fact, for either

$$\xi = a \quad \text{or} \quad \xi = x \quad , \tag{3.16}$$

we obtain a maximum value which is less than R_n for

$$a < \xi < x \quad . \tag{3.17}$$

Similarly, for either

$$\xi = a \quad \text{or} \quad \xi = x \quad , \tag{3.18}$$

we obtain a minimum value which is greater than R_n for the condition (3.17).

For notational simplicity, we make the following notational definitions.

Definition 3.1. $R_{n,a} = \frac{(x-a)^n}{n!} f^{(n)}(a)$.

Definition 3.2. $R_{n,x} = \frac{(x-a)^n}{n!} f^{(n)}(x)$.

It is clear that either

$$R_{n,a} \leq R_n \leq R_{n,x} \tag{3.19}$$

or

$$R_{n,x} \leq R_n \leq R_{n,a} \quad . \tag{3.20}$$

We are now in a position to define the closed interval which contains R_n .

Definition 3.3: $R_{n,i} = [\min(R_{n,a}, R_{n,x}), \max(R_{n,a}, R_{n,x})]$.

THEOREM 3.2. The closed interval, $R_{n,i}$, contains R_n .

PROOF: The proof is immediate. For Definitions 3.1, 3.2, and 3.3, together with (3.19) and (3.20), clearly show that we have constructed a closed interval which contains R_n . Q.E.D.

The closed interval which contains the remainder term (3.14) of the Maclaurin series can be constructed in the same manner as was the closed interval for Taylor's series, with $a = 0$. Thus, we have from Definition 3.3, with $a = 0$,

$$R_{n,i} = [\min(R_{n,o}, R_{n,x}), \max(R_{n,o}, R_{n,x})] \quad . \tag{3.21}$$

This of course is obvious, for the Maclaurin series is the Taylor series with $a = 0$ and the condition $o < \xi < x$.

Taylor's Series Interval Algorithm

If we modify the usual form of the Taylor series (3.3), to make use of the closed interval $R_{n,i}$, which by THEOREM 3.2 contains R_n , we have

$$f(x) \subset \sum_{j=0}^{n-1} \frac{1}{j!} (x - a)^j f^{(j)}(a) + R_{n,i} \quad . \tag{3.22}$$

When $f(x)$ is any real function which can be computed by Taylor's series, and when the n^{th} derivative is either increasing or decreasing monotonically between a and x , then the first n terms can be computed in the

usual manner and then added to the closed interval $R_{n,i}$. The sum of the first n terms will be a real number, and by Definition 2.6 we can add real numbers and closed intervals.

In actual practice, however, it is often necessary to start a computation with an inexact value of a parameter x, say, $x \pm \epsilon$. This problem can be taken care of for a series expansion and a result obtained with a precise error bound if we employ interval arithmetic.

Let us assume that f(x) is a real function and meets the conditions for being expanded by Taylor's theorem, and also meets the condition imposed upon the n^{th} derivative. Further, suppose x is not precisely known, but rather that we have $x \pm \epsilon$, e.g., e^x with $x = 1 \pm 0.1$, or $\sin x$ with $x = 0.25 \pm 0.001$. The usual procedure is to calculate the value of the function for x, without regard to the ambiguity, and then to do some analysis concerning the error. This often amounts to an educated guess, simply because of the difficulties encountered in the computation required by the analysis. What follows is a procedure for simultaneously doing the initial computation and the error analysis, thus obtaining a closed interval in which the exact value of the function is known to lie.

If x is a real number, so are $x + \epsilon$ and $x - \epsilon$. Clearly then, $f(x + \epsilon)$ and $f(x - \epsilon)$ can both be calculated by Taylor's series (3.3) or by (3.22), i.e., Taylor's series with a closed interval remainder $R_{n,i}$, assuming increasing or decreasing monotonicity of the n^{th} derivative between a and x.

For notational purposes, let us define the following.

Definition 3.4: $f(x - \epsilon) \subset \sum_{j=0}^{n-1} \frac{1}{j!} (x - \epsilon - a)^j f^{(j)}(a) + R_{n,\alpha}$.

Definition 3.5: $R_{n,\alpha} = R_{n,i}$ with x replaced by $x - \epsilon$, i.e.,

(3.23) $$R_{n,\alpha} = [\min(R_{n,a}, R_{n,x-\epsilon}), \max(R_{n,a}, R_{n,x-\epsilon})] .$$

Definition 3.5 is simply based on the fact that the condition on the remainder term is now

(3.24) $$a < \xi < x - \epsilon .$$

Definition 3.6: Let α be equal to the sum of the first n terms of Definition 3.4. That is,

(3.25) $$\alpha = \sum_{j=0}^{n-1} \frac{1}{j!} (x - \epsilon - a)^j f^{(j)}(a) .$$

We can now restate $f(x - \epsilon)$ as

(3.26) $$f(x - \epsilon) \subset \alpha + R_{n,\alpha}$$

or

(3.27) $$f(x - \epsilon) \subset [\alpha + (R_{n,\alpha})_{min}, \alpha + (R_{n,\alpha})_{max}] ,$$

where by $(R_{n,\alpha})_{min}$ and $(R_{n,\alpha})_{max}$ we mean the left and right ends of the closed interval $R_{n,\alpha}$, respectively.

For $f(x + \epsilon)$ the following definitions may be made.

Definition 3.7: $f(x + \epsilon) \subset \sum_{j=0}^{n-1} \frac{1}{j!} (x + \epsilon - a)^j f^{(j)}(a) + R_{n,\beta}$

Definition 3.8: $R_{n,\beta} = R_{n,i}$ with x replaced by $x + \epsilon$, i.e.,

(3.28) $$R_{n,\beta} = [\min(R_{n,a}, R_{n,x+\epsilon}), \max(R_{n,a}, R_{n,x+\epsilon})] .$$

Definition 3.8 is based on the fact that the condition on the remainder term is now

(3.29) $$a < \xi < x + \epsilon .$$

Definition 3.9: Let β be equal to the sum of the first n terms of Definition 3.7. That is,

$$\beta = \sum_{j=0}^{n-1} \frac{1}{j!} (x + \epsilon - a)^j f^{(j)}(a) \quad . \tag{3.30}$$

We can now restate $f(x + \epsilon)$ as

$$f(x + \epsilon) \subset \beta + R_{n,\beta} \tag{3.31}$$

or

$$f(x + \beta) \subset [\beta + (R_{n,\beta})_{min}, \beta + (R_{n,\beta})_{max}] \tag{3.32}$$

where by $(R_{n,\beta})_{min}$ and $(R_{n,\beta})_{max}$ we mean the left and right ends of the closed interval $R_{n,\beta}$, respectively.

Thus, we are in a position to state succinctly a certain algorithm for computing a closed interval which contains all values of the function $f(x \pm \epsilon)$, and which also contains the upper and lower error bounds.

ALGORITHM 3.1. If $f(x)$ meets the conditions of Taylor's theorem, and if the n^{th} derivative is monotonically increasing or decreasing between a and x, then for $f(x \pm \epsilon)$ we have the Taylor Series Interval Algorithm

$$f(x \pm \epsilon) \subset \left[\min\left(\alpha + (R_{n,\alpha})_{min}, \beta + (R_{n,\beta})_{min}\right), \max\left(\alpha + (R_{n,\alpha})_{max}, \beta + (R_{n,\beta})_{max}\right)\right] . \tag{3.33}$$

If in this section, a is everywhere replaced by zero, we then obtain the Maclaurin Series Interval Algorithm, which is identical in notation with (3.33), since a does not appear explicitly in that expression.

It should be noted that if $\epsilon = 0$, then α and β in (3.25) and (3.30) are identical and (3.33) becomes (3.22).

Example

It has been shown that by using interval arithmetic, we can obtain explicit closed intervals containing the exact solution and the error bounds for any function which can be calculated by the Taylor or Maclaurin series, and meets the condition imposed in the last section for the n^{th} derivative.

We shall demonstrate the use of the algorithm developed in the preceding section by using the Maclaurin series to calculate $\sin(x \pm \epsilon)$. It is assumed that x lies in the closed interval 0 to $\frac{\pi}{2}$. (This assumption is not really a restriction on the generality of the algorithm, as any value of x for the sine function can be scaled to be within the limits imposed.) It will be noted that the condition concerning the n^{th} derivative will always be met for $0 \leq x \leq \frac{\pi}{2}$.

No a priori analysis will be required to determine the number of terms needed to achieve any desired decimal place accuracy. If we simply calculate $f(x - \epsilon)$ and $f(x + \epsilon)$ according to Definitions 3.4 and 3.7 respectively, on a term by term basis, including the remainder for each term, we will then obtain closed intervals which contain $f(x - \epsilon)$ and $f(x + \epsilon)$. The closed intervals so found can have any desired accuracy. We have only to notice when the width of the remainder intervals, $R_{n,\alpha}$ and $R_{n,\beta}$, is less than the prescribed error. Once we have done this, we have the necessary elements for substituting in (3.33).

Remembering that the algorithm is planned for use on a digital computer, an experienced programmer can easily design a general purpose routine for any function which meets the required conditions.

We shall now illustrate these matters by finding a closed interval which contains $\sin(1 \pm 0.01)$ to five decimal place accuracy. The error bound will then be in the sixth place.

We first find $f(x - \epsilon)$ or $\sin(0.99)$ from (3.27). From (3.25), we have

$$\alpha = \sum_{j=0}^{n-1} \frac{1}{j!} (0.99)^j f^{(j)}(o) \quad . \tag{3.34}$$

For $R_{n,\alpha}$ we now make use of (3.23), noting, however, that since

$$f^{(n)}(x) = \sin\left(x + \frac{n\pi}{2}\right) , \tag{3.35}$$

we have

$$f^{(n)}(o) = \sin\left(\frac{n\pi}{2}\right) \quad . \tag{3.36}$$

Thus, since

$$0 < \xi < 0.99 \tag{3.37}$$

it is seen that

$$R_{n,o} = \frac{x^n}{n!} f^{(n)}(o) \tag{3.38}$$

and

$$R_{n,0.99} = \frac{x^n}{n!} f^{(n)}(0.99) \quad . \tag{3.39}$$

However, for the sine function restricted to the first quadrant, the minimum value of the function is zero and the maximum value is unity. Hence, for our problem, we can consider

$$R_{n,o} = 0 \tag{3.40}$$

and

$$R_{n,0.99} = (-1)^{n-1} \frac{x^n}{n!} \quad . \tag{3.41}$$

Relation (3.23) thus becomes

$$R_{n,\alpha} = \left[\min\left(0, (-1)^{n-1}\frac{x^n}{n!}\right), \max\left(0, (-1)^{n-1}\frac{x^n}{n!}\right)\right] . \tag{3.42}$$

Now calculating on a term by term basis until the width of $R_{n,\alpha}$ is less than five decimal places, we have

$$\begin{aligned}\sin(0.99) \subset\ & 0.9900000 - 0.1617165 + 0.0079249 \\ & - 0.0001849 + 0.0000025 \\ & + [-0.0000003, 0.0]\end{aligned} \tag{3.43}$$

or

$$\sin(0.99) \subset [0.836025, 0.836026] .$$

The exact value [3] of $\sin(0.99)$ to seven places is 0.8360259 .

We now find $f(x + \epsilon)$, i.e., $\sin(1.01)$ from (3.32). From (3.30) we have

$$\beta = \sum_{j=0}^{n-1} \frac{1}{j!}(1.01)^j f^{(j)}(o) . \tag{3.44}$$

For $R_{n,\beta}$ use is made of (3.28). Since

$$f^{(n)}(x) = \sin\left(x + \frac{n\pi}{2}\right) , \tag{3.45}$$

we have

$$f^{(n)}(o) = \sin\left(\frac{n\pi}{2}\right) . \tag{3.46}$$

And

$$0 < \xi < 1.01 ; \tag{3.47}$$

hence

$$R_{n,o} = \frac{x^n}{n!} f^{(n)}(o) \tag{3.48}$$

and

$$R_{n,1.01} = \frac{x^n}{n!} f^{(n)}(1.01) \ . \tag{3.49}$$

Again, making use of the fact that the function is always in the first quadrant, the minimum value is zero and the maximum value is unity. Thus

$$R_{n,o} = 0 \tag{3.50}$$

and

$$R_{n,1.01} = (-1)^{n-1} \frac{x^n}{n!} \ . \tag{3.51}$$

Hence, for (3.28) we have

$$R_{n,\beta} = \left[\min\left(0, (-1)^{n-1} \frac{x^n}{n!}\right), \max\left(0, (-1)^{n-1} \frac{x^n}{n!}\right)\right] . \tag{3.52}$$

Now if calculations are made until the width of $R_{n,\beta}$ is less than five decimal places, we have

$$\begin{aligned} \sin(1.01) \subset\ & 1.0100000 - 0.1717168 + 0.0087584 \\ & - 0.0002127 + 0.0000030 \\ & + [-0.0000003, 0] \end{aligned} \tag{3.53}$$

$$\sin(1.01) \subset [0.846831, 0.846832],$$

The exact value [4] of $\sin(1.01)$ to seven places is 0.8468318 .

We now have all the elements of (3.33) and the closed interval is

$$\sin(1 \pm 0.01) \subset [0.836025, 0.846832] \quad . \tag{3.54}$$

Clearly, all values of $\sin x$ for

$$0.99 \leq x \leq 1.01 \tag{3.55}$$

are included in the interval (3.54). For further computation, one can take the center point of the interval, keeping track of the error bounds; or one could do all further computation in intervals.

Round-off error, which has not been considered in our discussion, can be included in our intervals by simply widening them at one or both ends to include the full value of all numbers being computed. The interval arithmetic program listed in the Appendix is designed to do just this.

CHAPTER IV

APPLICATION TO THE INITIAL-VALUE PROBLEM OF FIRST-ORDER ORDINARY DIFFERENTIAL EQUATIONS

Introduction

Some familiar, preliminary concepts and theorems from calculus[1] will be presented first. We shall then prove an existence theorem for a first-order ordinary differential equation, and present a technique for finding a closed interval containing its solution.

The concept of uniform convergence is basic to the proof of the existence theorem. If $f_n(x)$, $(n = 1, 2, \ldots)$, is a sequence of functions of x, each defined in a closed interval $a \leq x \leq b$, the sequence is said to converge to $f(x)$ in the closed interval if for each x of the closed interval,

$$\lim_{n \to \infty} f_n(x) = f(x) \ . \tag{4.1}$$

If (4.1) holds, then for a fixed x, we may make $|f_n(x) - f(x)|$ as small as we please, simply by choosing n sufficiently large.

Definition 4.1: (Uniform Convergence [1]) The sequence $f_n(x)$ is said to converge uniformly to $f(x)$, for $a \leq x \leq b$, if for every $\epsilon > 0$ an integer N can be found, such that

$$|f_n(x) - f(x)| < \epsilon \qquad (n \geq N) \tag{4.2}$$

for all x.

N, of course, is dependent upon ϵ, but is independent of x.

---

[1] See, for example, references [2] and [4]. For completeness we reproduce some of the proofs.

THEOREM 4.1. (CONTINUITY OF THE LIMIT FUNCTION [2]). If the sequence $f_n(x)$ converges uniformly to $f(x)$ for $a \le x \le b$ and, for each n, $f_n(x)$ is continuous on $a \le x \le b$, then $f(x)$ is continuous on $a \le x \le b$.

PROOF: Let x_o be any point in the closed interval $a \le x \le b$. Then

$$f(x) - f(x_o) = \left(f(x) - f_n(x)\right) + \left(f_n(x) - f_n(x_o)\right) + \left(f_n(x_o) - f(x_o)\right) \quad , \tag{4.3}$$

and

$$|f(x) - f(x_o)| \le |f(x) - f_n(x)| + |f_n(x) - f_n(x_o)| + |f_n(x_o) - f(x_o)| \quad . \tag{4.4}$$

If $\epsilon > 0$, then, by Definition 4.1, we can choose n large enough so that

$$|f_n(x) - f(x)| < \frac{\epsilon}{3} \tag{4.5}$$

for every x in the interval. Thus equation (4.4) becomes

$$|f(x) - f(x_o)| < |f_n(x) - f_n(x_o)| + \frac{2}{3}\epsilon \quad . \tag{4.6}$$

Now n is fixed and $f_n(x)$ is continuous at x_o. Therefore, if x is sufficiently close to x_o, we have

$$|f_n(x) - f_n(x_o)| < \frac{\epsilon}{3} \, . \tag{4.7}$$

By (4.6) we thus have

$$|f(x) - f(x_o)| < \epsilon \quad , \tag{4.8}$$

which proves that the limit of a uniformly convergent sequence of continuous functions is continuous.

THEOREM 4.2. (INTEGRATION OF SEQUENCES [3]). If the sequence $f_n(x)$ converges uniformly to $f(x)$ on $a \le x \le b$ and, if for each n, $f_n(x)$ is continuous for $a \le x \le b$, then

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \int_a^b f_n(x)\,dx \quad . \tag{4.9}$$

PROOF: From the integral calculus, we have

$$\int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx = \int_a^b \left(f_n(x) - f(x)\right) dx \tag{4.10}$$

$$\left| \int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx \right| \le \int_a^b |f_n(x) - f(x)|\,dx \quad . \tag{4.11}$$

Let $\epsilon > 0$. Choose N, independent of x, sufficiently large enough so that for $N \le n$,

$$|f_n(x) - f(x)| < \frac{\epsilon}{b-a} \quad , \quad a \le x \le b \quad . \tag{4.12}$$

Equation (4.12) is true, since by hypothesis we have uniform convergence. Thus, for $N \le n$,

$$\int_a^b |f_n(x) - f(x)|\,dx < \int_a^b \frac{\epsilon}{b-a}\,dx = \epsilon \quad . \tag{4.13}$$

Hence

$$\left| \int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx \right| < \epsilon \quad . \tag{4.14}$$

But (4.14) implies that (4.9) is true, and the proof is complete.

THEOREM 4.3. (COMPARISON TEST FOR CONVERGENCE [4]). If $|a_n| \le b_n$ for $n = 1, 2, \ldots$ and if $\sum_{n=1}^{\infty} b_n$ converges, then $\sum_{n=1}^{\infty} a_n$ is absolutely convergent.

PROOF: Since for $0 \leq a_n$, $n = 1, 2, \ldots$, the series $\sum_{n=1}^{\infty} a_n$ is either convergent or properly divergent,[2] let us assume that the series $\sum |a_n|$ is properly divergent. Then

$$\lim_{n \to \infty} \sum_{i=1}^{n} |a_i| = \infty \quad . \tag{4.15}$$

But since $|a_n| \leq b_n$, we would have

$$\lim_{n \to \infty} \sum_{i=1}^{n} b_i = \infty \quad , \tag{4.16}$$

so that $\sum b_n$ would be divergent, which is contrary to the assumption. Therefore $\sum_{n=1}^{\infty} |a_n|$ converges and $\sum_{1}^{\infty} a_n$ is absolutely convergent.

THEOREM 4.4. (WEIERSTRASS M-TEST [5]). Let $\sum_{n=1}^{\infty} u_n(x)$ be a series of functions all defined for $a \leq x \leq b$. If there is a convergent series of constants $\sum_{n=1}^{\infty} M_n$, such that

$$|u_n(x)| \leq M_n \tag{4.17}$$

for all $a \leq x \leq b$, then the series $\sum_{n=1}^{\infty} u_n(x)$ converges absolutely for each x in the given interval and is uniformly convergent in the interval.

PROOF: Since for each term of the series $\sum_{n=1}^{\infty} u_n(x)$, we have

$$|u_n(x)| \leq M_n \, , \tag{4.18}$$

and $\sum_{n=1}^{\infty} M_n$ is convergent by hypothesis, then by THEOREM 4.3, the series $\sum_{n=1}^{\infty} u_n(x)$ is absolutely convergent. To show that the series is uniformly

---

[2]The series $\sum a_n$ is said to be properly divergent if, for the sequence S_n of partial sums, either $\lim S_n = +\infty$ or $\lim S_n = -\infty$.

convergent, let

$$S = \sum_{n=1}^{\infty} u_n(x) \tag{4.19}$$

and let

$$R_n = S - S_n \tag{4.20}$$

where S_n is the first n term of the series $\sum u_n(x)$. Hence R_n is the remainder after n terms of $\sum u_n(x)$. Then

$$|R_n(x)| = |u_{n+1}(x) + u_{n+2}(x) + \ldots| \leq |u_{n+1}(x)| + |u_{n+2}(x)| + \ldots \leq M_{n+1} + M_{n+2} + \ldots \quad . \tag{4.21}$$

If T_n denotes the remainder after n terms of the convergent series $\sum M_n$, then

$$T_n = M_{n+1} + M_{n+2} + \ldots \tag{4.22}$$

and

$$|R_n(x)| \leq T_n \quad . \tag{4.23}$$

Since $\sum M_n$ is a series of constants, for each given $\epsilon > 0$, some N can be found such that $T_n < \epsilon$ for $n \geq N$. Thus

$$|R_n(x)| \leq T_n < n \quad , \quad n \geq N \quad . \tag{4.24}$$

As N depends only on ϵ and not on x , we have the desired uniform convergence and the proof is complete.

Outline of the Picard Method

We are now in a position to prove the existence of a unique solution for the initial-value problem of a first-order ordinary differential equation. The proof is Picard's method of successive approximations. We first present two lemmas, and then the existence theorem.

LEMMA 4.1. Consider the functional sequence $y_1(x), y_2(x), \ldots, y_n(x)$ defined by

$$(4.25)\qquad \begin{cases} y_1(x) = y_o + \int_{x_o}^{x} f(t, y_o)\, dt \\ y_2(x) = y_o + \int_{x_o}^{x} f\big(t, y_1(t)\big)\, dt \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ y_n(x) = y_o + \int_{x_o}^{x} f\big(t, y_{n-1}(t)\big)\, dt \quad , \end{cases}$$

where (x_o, y_o) is a point within a rectangular domain D defined by

$$(4.26)\qquad |x - x_o| \le a \quad , \quad |y - y_o| \le b \quad .$$

Let $f(x,y)$ be a single-valued continuous function of x and y, and let M be the upper bound of $|f(x,y)|$ in D. Further, let h be the smaller of a and b/M. Now if x is in the closed interval

$$(4.27)\qquad x_o \le x \le x_o + h \quad ,$$

then

$$(4.28)\qquad |y_n(x) - y_o| \le b .$$

PROOF (by mathematical induction): Let x be within the closed interval (4.27) and let $y_n(x)$ be defined by (4.25). Then for $n = 1$,

$$|y_1(x) - y_o| \leq \int_{x_o}^{x} |f(t, y_o)| \, dt \quad . \tag{4.29}$$

By hypothesis M is the upper bound of $|f(x, y)|$ in D, hence

$$|y_1(x) - y_o| \leq M(x - x_o) \leq Mh \leq b \quad . \tag{4.30}$$

Therefore (4.28) is true for $n = 1$. Assume now that (4.28) is true for $n - 1$, that is, that

$$|y_{n-1}(x) - y_o| \leq b \quad . \tag{4.31}$$

Then it follows that

$$|f\left(t, y_{n-1}(t)\right)| \leq M \quad , \tag{4.32}$$

since M is the upper bound of $|f(x, y)|$ in D. Hence

$$\int_{x_o}^{x} |f\left(t, y_{n-1}(t)\right)| \, dt \leq M(x - x_o) \leq Mh \leq b \quad . \tag{4.33}$$

But from (4.25)

$$y_n(x) - y_o = \int_{x_o}^{x} f\left(t, y_{n-1}(t)\right) \, dt \quad . \tag{4.34}$$

Therefore, we have

$$|y_n(x) - y_o| \leq \int_{x_o}^{x} |f\left(t, y_{n-1}(t)\right)| \, dt \leq M(x - x_o) \leq Mh \leq b \quad , \tag{4.35}$$

and (4.28) holds for n. Thus (4.28) holds for all $n \geq 1$ and the lemma is proved.

LEMMA 4.2. Let the conditions of LEMMA 4.1 hold. Further, let K be the Lipschitz constant such that for any two points in D of the same abscissa, say (x, y_1) and (x, y_2),

$$|f(x, y_1) - f(x, y_2)| < K |(y_2 - y_1)| \quad . \tag{4.36}$$

Then

$$|y_n(x) - y_{n-1}(x)| < \frac{MK^{n-1}}{n!} \left|(x - x_o)^n\right| \quad . \tag{4.37}$$

PROOF (by mathematical induction): Let x be within the interval (4.27) and let $y_n(x)$ be defined by (4.25). Then for $n = 1$,

$$|y_1(x) - y_o(x)| < \frac{MK^o}{1!} \left|(x - x_o)^o\right| < M \quad . \tag{4.38}$$

Clearly, this is true. Assume now that (4.37) is true for $n - 1$. That is, that

$$|y_{n-1}(x) - y_{n-2}(x)| < \frac{MK^{n-2}}{(n-1)!} \left|(x - x_o)^{n-1}\right| \quad . \tag{4.39}$$

For n,

$$|y_n(x) - y_{n-1}(x)| \leq \int_{x_o}^{x} |f(t, y_{n-1}(t)) - f(t, y_{n-2}(t))| \, dt \tag{4.40}$$

from (4.25). Applying the Lipschitz condition gives

$$|y_n(x) - y_{n-1}(x)| < K \int_{x_o}^{x} |y_{n-1}(t) - y_{n-2}(t)| \, dt \, . \tag{4.41}$$

The integrand of (4.41) is the same as the left side of equation (4.39). Hence

$$(4.42) \qquad |y_n(x) - y_{n-1}(x)| < \frac{MK^{n-1}}{(n-1)!} \int_{x_o}^{x} \left|(t - x_o)^{n-1}\right| \, dt$$

$$< \frac{MK^{n-1}}{n!} \left|(x - x_o)^n\right| \quad .$$

Thus (4.37) holds for all $n \geq 1$ and the proof is complete.

THEOREM 4.5. (EXISTENCE THEOREM FOR $y' = f(x,y)$ [6]). Let the conditions of LEMMAS 4.1 and 4.2 hold. Then there exists a unique continuous function of x, say $f(x)$, for

$$(4.43) \qquad x_o \leq x \leq x_o + h \quad ,$$

which satisfies the differential equation

$$(4.44) \qquad y' = f(x,y)$$

with initial conditions

$$(4.45) \qquad y(x_o) = y_o$$

and which reduces to y_o when $x = x_o$.

PROOF: Let x be within the interval (4.43) and consider the sequence of functions

$$(4.46) \qquad y_1(x), y_2(x), \ldots, y_n(x)$$

defined by (4.25). Then by LEMMA 4.1

$$(4.47) \qquad |y_n(x) - y_o| \leq b \quad .$$

It follows that

$$(4.48) \qquad f\big(x, y_n(x)\big) \leq M$$

and therefore the $f_n(x)$ are well defined. By LEMMA 4.2 and THEOREM 4.4 it is clear that the series

$$y_o + \sum_{r=1}^{\infty} \left(y_r(x) - y_{r-1}(x)\right) \tag{4.49}$$

is absolutely and uniformly convergent when x is in the interval (4.43), for the series $\sum M_n$ converges by the ratio test.[3] Also, each term of (4.49) is a continuous function of x.

Now

$$y_n(x) = y_o + \sum_{r=1}^{n} \left(y_r(x) - y_{r-1}(x)\right) \quad . \tag{4.50}$$

Therefore, by THEOREM 4.1, the limit function

$$y(x) = \lim_{n\to\infty} y_n(x) \tag{4.51}$$

exists and is a continuous function of x in the interval (4.43).

To show that the limit function $y(x)$ satisfies the differential equation, we note that the sequence

$$g_n(x) = f\left(x, y_n(x)\right) \tag{4.52}$$

converges uniformly to $f\left(x, y(x)\right)$ for $x_o \le x \le x_o + h$. This, of course, follows from the uniform convergence of $y_n(x)$, for by the Lipschitz condition (4.36),

---

[3]The ratio test for convergence requires that given a series of positive terms $\sum_{n=0}^{\infty} a_n$, $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = r < 1$.

$$|f(x,y(x)) - f(x,y_n(x))| \leq K\,|y(x) - y_n(x)| \quad . \tag{4.53}$$

Thus, by THEOREM 4.2,

$$\begin{cases} y(x) = \lim\limits_{n\to\infty} y_n(x) \\ \\ \quad = \lim\limits_{n\to\infty} \left[\int\limits_{x_o}^{x} f(t, y_{n-1}(t))\, dt + y_o \right] \\ \\ \quad = \int\limits_{x_o}^{x} f(t,y(t))\, dt + y_o \quad . \end{cases} \tag{4.54}$$

The function $f(t,y(t))$ is continuous in the interval $x_o \leq x \leq x_o + h$, and therefore

$$y'(x) = \frac{d}{dx} \int\limits_{x_o}^{x} f(t,y(t))\, dt = f(x,y(x)) \tag{4.55}$$

which shows that the limit function $y(x)$ satisfies the differential equation and reduces to y_o when $x = x_o$.

We will now show that the solution $y(x)$ is unique. Let $\bar{y}(x)$ be a solution to the differential equation, different from $y(x)$, and satisfying the initial condition $\bar{y}(x_o) = y_o$. Let $\bar{y}(x)$ be continuous in the interval $(x_o, x_o + h')$ where $h' \leq h$ and h' is such that the condition

$$|\bar{y}(x) - y_o| < b \tag{4.56}$$

holds. Since $\bar{y}(x)$ is a solution, it satisfies the integral equation

$$\bar{y}(x) = y_o + \int\limits_{x_o}^{x} f(t,\bar{y}(t))\, dt \tag{4.57}$$

and thus

$$\bar{y}(x) - y_n(x) = \int_{x_o}^{x} \left[f\left(t, \bar{y}(t)\right) - f\left(t, y_{n-1}(t)\right) \right] dt \quad . \tag{4.58}$$

For $n = 1$,

$$\bar{y}(x) - y_1(x) = \int_{x_o}^{x} \left[f\left(t, \bar{y}(t)\right) - f(t, y_o) \right] dt \quad , \tag{4.59}$$

and from the Lipschitz condition we have

$$|\bar{y}(x) - y_1(x)| < Kb(x - x_o) \quad . \tag{4.60}$$

For $n - 1$, assume

$$|\bar{y}(x) - y_{n-1}(x)| < \frac{1}{(n-1)!} K^{n-1} b(x - x_o)^{n-1} \quad . \tag{4.61}$$

Then for n, we have

$$\left.\begin{aligned} |\bar{y}(x) - y_n(x)| &< \left| \int_{x_o}^{x} \left[f\left(t, \bar{y}(t)\right) - f\left(t, y_{n-1}(t)\right) \right] dt \right. \\ &< K^{n-1} \int_{x_o}^{x} |\bar{y}(t) - y_{n-1}(t)| \, dt \\ &< \frac{K^{n-1}}{(n-1)!} \int_{x_o}^{x} Kb(t - x_o)^{n-1} \, dt \\ &< \frac{1}{(n)!} K^n b(x - x_o)^n \quad . \end{aligned}\right\} \tag{4.62}$$

Hence

$$\bar{y}(x) = \lim_{n \to \infty} y_n(x) \tag{4.63}$$

but from (4.51)

$$y(x) = \lim_{n \to \infty} y_n(x) \quad . \tag{4.64}$$

Therefore

$$y(x) = \bar{y}(x) \tag{4.65}$$

for all x in the interval $(x_o , x_o + h')$ and the two solutions are identical. Thus there is one and only one continuous solution of the differential equation which satisfies the initial conditions. This completes the proof.

Interval Integrals

This section will be concerned with certain definitions, theorems and notational conventions which we shall need for a discussion of the interval arithmetic approach to the initial-value problem of first-order ordinary differential equations.

A rational interval expression has been given by Definition 2.17.

Definition 4.2: A regular domain [7] is a set D of n-tuples of intervals where

$$([a_1 , a_2] , [a_3 , a_4] , \ldots , [a_{n-1} , a_n]) \in D \tag{4.66}$$

and

$$[a'_1 , a'_2] \subset [a_1 , a_2] , \ldots , [a'_{n-1} , a'_n] \subset [a_{n-1} , a_n] \tag{4.67}$$

imply that

$$([a'_1 , a'_2] , [a'_3 , a'_4] , \ldots , [a'_{n-1} , a'_n]) \in D \quad . \tag{4.68}$$

We will denote the Cartesian product by the symbol $\textcircled{X}$.

The set of all possible closed intervals contained within a given closed interval $[a_1, a_2]$ will be denoted by $S_{[a_1, a_2]}$. Thus

$$S_{[a_1, a_2]} = \{[a_1', a_2'] \mid [a_1', a_2'] \subset [a_1, a_2]\} \quad . \tag{4.69}$$

A regular domain is a union of sets of the form:

$$S_{[a_1, a_2]} \otimes S_{[a_3, a_4]} \otimes \cdots \otimes S_{[a_{n-1}, a_n]} \quad . \tag{4.70}$$

Definition 4.3: A rational interval function [8], F, is defined as the mapping

$$f: \quad D \rightarrow S \tag{4.71}$$

with regular domain $D \subset S^n$, where S is the set of all closed intervals.

Thus if f is the real restriction of F, where F is a rational interval function with domain D, then

$$\cup f(x_1, x_2, \ldots, x_n) \subset F([x_1, x_2], [x_3, x_4], \ldots, [x_{2n-1}, x_{2n}]) , \tag{4.72}$$

where $x_i \in [x_{2i-1}, x_{2i}]$, $i = 1, 2, \ldots, n$.

For a single variable, if the real function f is the real restriction of a rational interval function F, with regular domain D, then

$$f(x) \subset F[x_1, x_2] \quad , \quad x \in [x_1, x_2] \quad . \tag{4.73}$$

If $x_1 = x_2$, i.e., F is restricted to the domain S* of Definition 2.5, then for the real restriction of F, we have

$$f(x) = F[x_1, x_1] \quad , \quad x = x_1 \tag{4.74}$$

Definition 4.4: The width of the closed interval [a,b] will be defined by

$$w([a,b]) = b - a \quad . \tag{4.75}$$

Analogous to the continuity of the points on the real line, closed intervals are continuous.

Definition 4.5: The distance function P :

$$P([a,b],[c,d]) = \max(|a-c|, |b-d|) \quad . \tag{4.76}$$

Moore [9] proves that P is a metric[4] on S. Clearly,

$$\lim_{P \to 0} ([a,b] - [c,d]) \to [0,0] \equiv 0 \quad . \tag{4.77}$$

The following four theorems concerning rational interval functions have been proved by Moore [10] and are stated here without proof:

THEOREM 4.6. Rational interval functions are continuous.

THEOREM 4.7. There exists a positive real number K, independent of the method of subdivision of the interval $[x_1, x_{2n}]$, such that

$$\bigcup_{i=1}^{n} F([x_{i-1}, x_i]) = \bigcup_{x=[x_o, x_n]} f(x) + [\epsilon', \epsilon''] \tag{4.78}$$

with

$$0 \in [\epsilon', \epsilon''] \tag{4.79}$$

and

$$w([\epsilon', \epsilon'']) \leq K \max w([x_{i-1}, x_i]) \quad . \tag{4.80}$$

---

[4] A metric is a single-valued, positive real function $\rho(x,y)$ satisfying the conditions: (1) $\rho(x,y) = 0$ if and only if $x = y$; (2) $\rho(x,y) = \rho(y,x)$; and (3) $\rho(x,y) + \rho(y,z) \geq \rho(x,z)$.

THEOREM 4.8. (LIPSCHITZ CONDITION FOR INTERVALS.) For any rational interval function with regular domain $S_{[a_1, a_2]}$ and real valued real restriction, there exists a real number K such that

$$[x_1, x_2] \subset [a_1, a_2] \tag{4.81}$$

implies that

$$w\left(F([x_1, x_2])\right) \leq Kw([x_1, x_2]) \quad . \tag{4.82}$$

The following theorem states that if f(x) is a rational function, it can be integrated by the interval extension $F([x_o, x_n])$, for $x \in [x_o, x_n]$ with strict error bounds maintained.

THEOREM 4.9.

$$\sum_{i=1}^{n} F([x_{i-1}, x_i])\, w([x_{i-1}, x_i]) = \int_a^x f(x)\, dx + [\epsilon', \epsilon''] \tag{4.83}$$

with

$$0 \in [\epsilon', \epsilon''] \tag{4.84}$$

and

$$w([\epsilon', \epsilon'']) \leq (x - a)\, K \max w([x_{i-1}, x_i]) \quad , \tag{4.85}$$

where K is the Lipschitz constant.

An Interval Approach to the First-Order Problem [11]

A method shall now be presented for finding a closed interval which contains the solution to the first-order differential equation

$$y' = f(x, y) \tag{4.86}$$

with initial conditions

$$y(x_o) = y_o \,. \tag{4.87}$$

It has been shown in THEOREM 4.5, that when a real valued function f is continuous on a region in the xy-plane, say

$$D_f = [x_o, a] \otimes [b_1, b_2] \tag{4.88}$$

with $x_o < a$ and $b_1 \leq y_o \leq b_2$ and that when f satisfies a Lipschitz condition on D_f

$$|f(x,y_1) - f(x,y_2)| \leq K\,|y_1 - y_2| \tag{4.89}$$

for some positive real number K, then there exists exactly one solution to (4.86) and (4.87) in $[x_o, x^*]$ for x^* such that for all $(x,y) \in D$

$$y_o + (x^* - x_o)\, f(x,y) \in [b_1, b_2] \quad . \tag{4.90}$$

It will be assumed, hereafter, that F is a rational interval valued function, as defined by Definition 4.3, on the regular domain

$$D_F = S_{[x_o, a]} \otimes S_{[b_1, b_2]} \tag{4.91}$$

and that F satisfies the following conditions:

(1) F is continuous;

(2) F is restricted to the domain

$$D_f = [x_o, a] \otimes [b_1, b_2] \quad , \tag{4.92}$$

where f is a real valued function;

(3) F is monotonic inclusive, i.e. if

$$\left[x_1', x_2'\right] \subset [x_1, x_2], \left[y_1', y_2'\right] \subset [y_1, y_2] \quad , \tag{4.93}$$

then

$$F\left(\left[x_1', x_2'\right], \left[y_1', y_2'\right]\right) \subset F([x_1, x_2], [y_1, y_2]) \quad ; \tag{4.94}$$

(4) There is a real number K_F such that

$$w\left(F([x_1, x_2], [y_1, y_2])\right) \leq K_F \max\left(w([x_1, x_2]), w([y_1, y_2])\right) \quad . \tag{4.95}$$

Note that if F is a rational interval function on D_F with real restriction f on D_f, then the above four conditions are satisfied by F.

The foregoing conditions (1), (2), (3), (4) imply that the Lipschitz condition (4.89) is met. For, assume

$$w\left(F([x_1, x_2], [y_1, y_2])\right) \leq K_F \max\left(w([x_1, x_2]), w([y_1, y_2])\right) \quad . \tag{4.96}$$

Then

$$w\left(F([x,x], [y_1, y_2])\right) \leq K_F\, w([y_1, y_2]) \tag{4.97}$$

for real $x_o \leq x \leq a$. Since f is the real restriction of F, by hypothesis, we have

$$f(x,y) \in F([x,x], [y_1, y_2]) \tag{4.98}$$

whenever

$$y \in [y_1, y_2] \quad . \tag{4.99}$$

Thus

$$f(x,y_1) - f(x,y_2) \in F([x,x], [y_1, y_2]) - F([x,x], [y_1, y_2]) \quad . \tag{4.100}$$

Now

$$[a,b] - [a,b] = [-1,1]\, w([a,b]) \quad , \tag{4.101}$$

so that

$$|f(x,y_1) - f(x,y_2)| \leq w\Big(F([x,x]\,,\,[y_1,y_2])\Big) \leq K_F\,|y_1 - y_2| \quad , \tag{4.102}$$

and hence

$$|f(x,y_1) - f(x,y_2)| = \max\,|f(x,y_1) - f(x,y_2)| \leq K_F\,|y_1 - y_2| \quad . \tag{4.103}$$

Since K_F serves as a Lipschitz constant for F, we conclude that conditions (1), (2), (3), (4) guarantee the existence and uniqueness in the closed interval $[x_o, x^*]$ of a solution to (4.86), (4.87) whenever f is the real restriction of F.

If

$$y_o \in [y_1, y_2] \subseteq [b_1, b_2] \quad , \tag{4.104}$$

then the equation

$$[y_1, y_2] + ([x_1^*, x_2^*] - [x_o, x_o])\, F(D_f) = [b_1, b_2] \tag{4.105}$$

has a solution

$$[x_1^*, x_2^*] \quad , \quad \text{where} \quad w([x_1^*, x_2^*]) > 0 \quad . \tag{4.106}$$

By $F(D_f)$ we mean the range of values of slopes of the interval function F over the region in the xy-plane D_f.

$$F(D_f) = F([x_o, a]\,,\,[b_1, b_2]) \quad . \tag{4.107}$$

We define

$$[x_o, x^*] = [x_o, a] \cap [x_1^*, x_2^*] \quad . \tag{4.108}$$

Thus we can compute a closed interval, $[x_o, x^*]$, in which existence and uniqueness of a solution to (4.86), (4.87) is guaranteed.

The method of computing the closed interval $[y_1, y_2]$ which contains the solution y is as follows. Let

$$\left[x_{i-1}^{(n)}, x_i^{(n)}\right] = [x_o, x_o] + [i-1, i] \frac{w([x_o, x^*])}{[n, n]} \quad , \quad (i = 1, 2, \ldots, n) \quad . \tag{4.109}$$

$$\left[b_{i1}^{(n)}, b_{i2}^{(n)}\right] = \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + [0, 1] \frac{w([x_o, x^*])}{[n, n]} F(D_f) \quad , \quad (i = 1, 2, \ldots, n) \quad . \tag{4.110}$$

$$\left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right] = \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + \frac{w([x_o, x^*])}{[n, n]} F\left(\left[x_{i-1}^{(n)}, x_i^{(n)}\right], \left[b_{i1}^{(n)}, b_{i2}^{(n)}\right]\right) \quad , \quad (i = 1, 2, \ldots, n) \quad . \tag{4.111}$$

The i's in formulas (4.109), (4.110) and (4.111) refer to the subdivision of the interval $[x_o, x^*]$ into n subintervals. Formula (4.111) may be written in the following form:

$$\left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right] = \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + \frac{w([x_o, x^*])}{[n, n]} F\left(\left[x_{i-1}^{(n)}, x_i^{(n)}\right], \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + [0, 1] \frac{w([x_o, x^*])}{[n, n]} F(D_f)\right) \quad , \quad (i = 1, 2, \ldots, n) \quad . \tag{4.112}$$

The recursive relation (4.112) gives the interval method for the solution of a first-order ordinary differential equation in its simplest form.

The solution to (4.86), (4.87), i.e. y, satisfies

$$y(x) \in y\left(\left[x_{i-1}^{(n)}, x_{i-1}^{(n)}\right]\right) + [0,1]\frac{w[x_o, x^*]}{[n,n]} F(D_f) \tag{4.113}$$

for

$$x \in \left[x_{i-1}^{(n)}, x_i^{(n)}\right] , \tag{4.114}$$

and if

$$y\left(\left[x_{i-1}^{(n)}, x_{i-1}^{(n)}\right]\right) \in \left[y_{i-1}^{(n)}, y_{i-1}^{(n)}\right]$$

then

$$y(x) = y\left(\left[x_{i-1}^{(n)}, x_{i-1}^{(n)}\right]\right) + \int_{x_{i-1}}^{x} f\left(x', y(x')\right) dx \quad , \quad x \in \left[x_{i-1}^{(n)}, x_i^{(n)}\right] \tag{4.115}$$

so that

$$y(x) \in \left[y_{i-1,1}^{(n)}, y_{i-1,1}^{(n)}\right] + \left(x - x_{i-1}^{(n)}\right) F\left(\left[x_{i-1}^{(n)}, x_i^{(n)}\right] , \left[y_{i-1,1}^{(n)} y_{i-1,2}^{(n)}\right] + [0,1]\frac{w([x_o, x^*])}{[n,n]} F(D_f)\right) \tag{4.116}$$

whenever $x \in \left[x_{i-1}^{(n)}, x_i^{(n)}\right]$. Also, if

$$w\left(\left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right]\right) = \max w\left(\left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right]\right) \tag{4.117}$$

then

$$w\left(\left[y_{i,1}^{(n)}, y_{1,2}^{(n)}\right]\right) \le w\left(\left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right]\right) + hK_F \max\left(h , w\left(\left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right]\right) + ch\right) \tag{4.118}$$

where

$$h = \frac{w[x_o, x^*]}{[n,n]} \tag{4.119}$$

and

$$c = w\left([0,1]\ F(D_f)\right) \quad . \tag{4.120}$$

Now we must "solve" the inequality (4.118). Let us first simplify the notation by setting

$$\begin{cases} w_i = w\left(\left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right]\right) \\ w_{i-1} = w\left(\left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right]\right) \end{cases} . \tag{4.121}$$

Thus we want to solve the inequality

$$w_i \leq w_{i-1} + hK_F \max(h, w_{i-1} + ch) \tag{4.122}$$

for w_i, i.e. eliminate the w's from the right hand side of (4.122) and obtain upper bounds on w_i in terms of the "step size"

$$h = \frac{x^* - x^o}{n} \quad . \tag{4.123}$$

We know that $h > 0$, so

$$\frac{w_i}{h} \leq \frac{w_{i-1}}{h} + hK_F \max\left(1, \frac{w_{i-1}}{h} + c\right) \quad ; \tag{4.124}$$

that is

$$\begin{cases} \frac{w_i}{h} \leq \frac{w_{i-1}}{h} + hK_F \\ \frac{w_i}{h} \leq \frac{w_{i-1}}{h}(1 + hK_F) + (hK_F)\, c \end{cases} . \tag{4.125}$$

Therefore

$$\frac{w_i}{h} \leq \frac{w_{i-1}}{h}(1 + hK_F) + (hK_F)\max(c, 1) \quad , \tag{4.126}$$

which is of the form

$$P_i \leq P_{i-1}(1 + q) + q_m \tag{4.127}$$

or

$$P_i \leq qm + (1 + q)\, P_{i-1} \tag{4.128}$$

where $P_i = \frac{w_i}{h}$, $P_{i-1} = \frac{w_{i-1}}{h}$,

$q = hK_F$, $m = \max(c, 1)$.

Hence

$$P_i \leq q_m + (1 + q)(qm + (1 + q)(qm + \ldots(1 + q)\, P_o))) \tag{4.129}$$

and

$$P_i \leq (qm)\left(1 + (1 + q) + (1 + q)^2 + \ldots + (1 + q)^{i-1}\right) + (1 + q)^i\, P_o \tag{4.130}$$

or

$$P_i \leq (qm)\left(\frac{(1 + q)^i - 1}{(1 + q) - 1}\right) + (1 + q)^i\, P_o \tag{4.131}$$

and

$$P_i \leq (m)\left((1 + q)^i - 1\right) + (1 + q)^i\, P_o \quad . \tag{4.132}$$

But here,

$$P_o = \frac{w_o}{h} = \frac{w([y_{0,1}, y_{0,2}])}{h} = 0 \tag{4.133}$$

so that

$$w\left(\left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right]\right) = hP_i \leq h \max(c, 1)\left((1 + hK_F)^i - 1\right) \tag{4.134}$$

and

$$(1 + hK_F)^i = \left((1 + hK_F)^{1/h}\right)^{(x_i - x_o)} , \tag{4.135}$$

since

$$ih = x_i - x_o . \tag{4.136}$$

It is fairly well known that

$$(1 + hx)^{1/h} < e^x , \quad h > 0 , \tag{4.137}$$

so

$$(1 + hK_K)^i < e^{K_F(x_i - x_o)} \tag{4.138}$$

and therefore

$$w\left(\left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right]\right) \leq h \max(c, 1)\left(e^{K_F(x_i - x_o)} - 1\right) . \tag{4.139}$$

This proves that at a fixed value for x_i, say $x_i = \alpha$, we can make

$$w\left(\left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right]\right) \to 0 \tag{4.140}$$

as $h \to 0$.

We define for $n = 1, 2, \ldots,$ the functions $y^{(n)}$ for all

$$x \in [x_o, x^*] \quad , \tag{4.141}$$

noting that

$$[x_o, x^*] = \bigcup_{i=1}^{n} \left[x_{i-1}^{(n)}, x_i^{(n)}\right] \tag{4.142}$$

and define $y^{(n)}(x)$ for

$$x \in \left[x_{i-1}^{(n)}, x_i^{(n)}\right] \quad . \tag{4.143}$$

Thus

$$y^{(n)}(x) \subseteq \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + \left(x - x_{i-1}^{(n)}\right) F\left(\left[x_{i-1}, x_i\right], \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + \left(\left[x_{i-1}^{(n)}, x_i^{(n)}\right] - \left[x_{i-1}^{(n)}, x_{i-1}^{(n)}\right]\right) F(D_f)\right) \quad . \tag{4.144}$$

The functions $y^{(n)}(x)$ are well defined since at $x_i^{(n)}$, the common end point of $\left[x_{i-1}^{(n)}, x_i^{(n)}\right]$ and $\left[x_i^{(n)}, x_{i+1}^{(n)}\right]$ we have

$$\left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right] = \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + \left(x_i^{(n)} - x_{i-1}^{(n)}\right) F\left(\left[x_{i-1}, x_i\right], \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + \left(\left[x_{i-1}^{(n)}, x_i^{(n)}\right] - \left[x_{i-1}^{(n)}, x_{i-1}^{(n)}\right]\right) F(D_f)\right) \quad . \tag{4.145}$$

The functions defined by (4.144) are continuous interval-valued functions and are "piecewise linear" in x. That is, for $0 \le t \le 1$, we have

$$y^{(n)}\left((1 - t)\, x_{i-1} + t x_i\right) = (1 - t)\left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + t[y_{i,1}, y_{i,2}] \quad . \tag{4.146}$$

Thus, it has been shown that the interval-valued function defined by (4.109), (4.110), (4.111), and (4.144) contain the corresponding components of the solution to (4.86) with (4.87).

Thus

(4.147) $$y(x) \in y^{(n)}(x) \quad \text{for} \quad x \in [x_o, x^*]$$

and the sequence of interval valued functions $y^{(1)}(x)$, $y^{(2)}(x)$, $y^{(3)}, \ldots$ converges uniformly to $y(x)$ for $x \in [x_o, x^*]$. Further, there is a real number K such that

(4.148) $$\max w\left(y^{(n)}(x)\right) \leq \frac{K}{n} \quad , \quad x \in [x_o, x^*] \quad .$$

Example

We consider the simple first-order differential equation

(4.149) $$y' = x + y$$

with initial conditions

(4.150) $$y_o = x_o = 0 \quad .$$

Our method will be to solve equation (4.112) for the closed interval $\left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right]$ which contains the exact solution of (4.149) under (4.150).

To solve equation (4.112), however, we must first determine x^*, which is done by using equations (4.104), (4.105), and (4.108). Let

(4.151) $$y_o = 0 \in [y_1, y_2] = [0,0] \subseteq [b_1, b_2] = [-1,1] \quad ,$$

and choose $a = 1$, and $x_o = 0$ as given by the initial conditions (4.150). From (4.107) we have

(4.152) $$F(D_f) = ([0,1], [-1,1]) \quad .$$

It should be noted that if we let $X = [x_o, a]$ and $Y = [b_1, b_2]$, then $F(D_f)$ can be written in either of the following forms

(4.153) $$F(D_f) = F(X, Y) = F(X \circledast Y)$$

and for our problem

(4.154) $$F(D_f) = X + Y = [0,1] + [-1,1] = [-1,2]$$

Substituting these values into equation (4.105) we have

(4.155) $$[0,0] + ([x_1^*, x_2^*] - [0.0])([-1,2]) = [-1,1]$$

or

(4.156) $$[x_1^*, x_2^*][-1,2] = [-1,1] \quad .$$

A solution to (4.156) is guaranteed by THEOREM 2.17. Hence, solving (4.156) gives us

(4.157) $$[x_1^*, x_2^*] = \left[-\frac{1}{2}, \frac{1}{2}\right] \quad .$$

We find x^* from (4.103),

(4.158) $$[x_o, x^*] = [0,1] \cap \left[-\frac{1}{2}, \frac{1}{2}\right] = \left[0, \frac{1}{2}\right]$$

and so

(4.159) $$x^* = \frac{1}{2} \quad .$$

We are now prepared to find the closed interval $\left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right]$ of equation (4.112) which contains the solution to (4.150) with (4.151). Thus

$$(4.160)\quad \left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right] = \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + \frac{w\left(\left[0, \frac{1}{2}\right]\right)}{[n,n]} \left([0,0] + [i-1, i] \frac{w\left[0, \frac{1}{2}\right]}{[n,n]} \right.$$

$$\left. + \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + [0,1] \frac{w\left[0, \frac{1}{2}\right]}{[n,n]} [-1, 2] \right)$$

$$= \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + \left[\frac{1}{2n}, \frac{1}{2n}\right] \left(\left[\frac{i-1}{2n}, \frac{i}{2n}\right] \right.$$

$$+ \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + \left[\frac{-1}{2n}, \frac{2}{2n}\right]$$

$$= \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] + \left[\frac{1}{2n}, \frac{1}{2n}\right] \left[y_{i-1}^{(n)}, y_{i-1,2}^{(n)}\right] + \left[\frac{i-2}{4n^2}, \frac{i+2}{4n^2}\right]$$

$$= \left[y_{i-1,1}^{(n)}, y_{i-1,2}^{(n)}\right] \left(1 + \frac{1}{2n}\right) + \left[\frac{i-2}{4n^2}, \frac{i+2}{4n^2}\right] ,$$

$$\text{for} \quad i = 1, 2, \ldots, n \quad \text{and} \quad y_o^{(n)} = 0 \quad .$$

Now the function $y(x) \subset \left[y_{i,1}^{(n)}, y_{i,2}^{(n)}\right]$ can be found by computing (4.160) for $i = 1, 2, \ldots, n$.

For $n = 10$, we have

$$(4.161)\qquad \left[y_{10,1}^{(10)}, y_{10,2}^{(10)}\right] = [0.097, 0.223] \quad .$$

Since the exact solution for the differential equation given in (4.149) is

$$(4.162)\qquad y = -x - 1 + e^x \quad ,$$

for $x = \frac{1}{2}$ we have

$$(4.163)\qquad y\left(\frac{1}{2}\right) \cong 0.149$$

and

$$(4.164)\qquad y\left(\frac{1}{2}\right) \subset \left[y_{10,1}^{(10)}, y_{10,2}^{(10)}\right] \quad .$$

Having computed $\left[y_{10,1}^{(10)}, y_{10,2}^{(10)}\right]$, we can select $\left[y_{10,1}^{(10)}, y_{10,2}^{(10)}\right]$ and $x^* = \frac{1}{2}$ as the new initial conditions, select a new $a > x^*$ and another $[b_1, b_2] \supset \left[y_{10,1}^{(10)}, y_{10,2}^{(10)}\right]$ and repeat the procedure. In this way we can solve the differential equation (4.149) with initial conditions (4.150).

To find a solution at a value, say x' where $x_o < x' < x^*$, we simply set $x^* = x'$ in equation (4.159). Thus a solution can be obtained for any given value of x.

REFERENCES

REFERENCES

Chapter I

(1) Paul S. Dwyer, Linear Computations, New York, 1951, pp. 11–25.

(2) Saul Gorn, "The Automatic Analysis and Control of Computing Errors," Journal of the Society of Industrial and Applied Mathematics, Vol. 2, No. 2, (June, 1954), pp. 69–81.

(3) R. E. Moore, Automatic Error Analysis in Digital Computation, LMSD-48421, Lockheed Missiles & Space Company, Sunnyvale, California, 1959, pp. 43–56.

(4) George Collins, Interval Arithmetic for Automatic Error Analysis, International Business Machines Corp., Mathematics and Applications Dept., New York, 1960, pp. 1–11.

(5) R. E. Moore, Interval Arithmetic and Automatic Error Analysis in Digital Computing, Technical Report No. 25, Stanford Univ., Stanford, California, 1962, p. 130.

(6) G. V. Gendzhoian, "On the Two-Sided Chaplygin Approximation of the Solution of Two-Point Boundary Problems," Izv. AN Arm SSR, Fiz. Matematika Nauk, Vol. 17, No. 3, 1964, pp. 21–26.

(7) R. G. Aliev, "On the Question of Specific Criteria for Estimates of the Lengths of Subcritical Intervals," Izv. VUZ, Matematika, No. 5 (42), 1964, pp. 3–7.

(8) R. G. Aliev, V. V. Ostroumov and S. A. Pak, "On Certain Properties of Cauchy Functions," Izv. VUZ, Matematika, No. 4 (41), 1964, pp. 9–11.

(9) S. A. Pak and E. S. Chichkin, "On the Existence of Upper and Lower Solutions of the Cauchy Problem of a Second Order Differential Equation," Izv. VUZ, Matematika, No. 5 (42), 1964, pp. 91–94.

Chapter II

(1) R. E. Moore, Interval Arithmetic and Automatic Error Analysis in Digital Computing, Technical Report No. 25, Stanford Univ., Stanford, California, 1962, p. 3.

(2) Ibid., p. 7.

(3) Ibid., p. 6.

(4) R. E. Moore, Automatic Error Analysis in Digital Computing, LMSD-48421, Lockheed Missiles & Space Company, Sunnyvale, California, 1959, p. 49.

(5) R. E. Moore, Interval Arithmetic and Automatic Error Analysis in Digital Computing, Technical Report No. 25, Stanford Univ., Stanford, California, 1962, pp. 7–8.

(6) L. P. Eisenhart, Continuous Groups of Transformations, Princeton, New Jersey, 1933, p. 15.

(7) E. Hille and R. S. Phillips, Functional Analysis and Semi-Groups, Providence, Rhode Island, 1957, pp. 256–257.

Chapter III

(1) A. E. Taylor, Advanced Calculus, Boston, 1955, pp. 110–117.

(2) F. B. Hildebrand, Introduction to Numerical Analysis, New York, 1956, pp. 22–24.

(3) M. Abramowitz and I. A. Stegum (ed.), Handbook of Mathematical Functions, AMS 55, Washington, D. C., 1964, p. 161.

(4) Ibid., p. 162.

Chapter IV

(1) Wilfred Kaplan, Ordinary Differential Equations, Reading, Massachusetts, 1958, p. 471.

(2) Angus E. Taylor, Advanced Calculus, Boston, 1955, p. 598.

(3) Ibid., pp. 599–600.

(4) Wilfred Kaplan, Advanced Calculus, Cambridge, Massachusetts, 1953, p. 314.

(5) Ibid., pp. 342–343.

(6) E. L. Ince, Ordinary Differential Equations, London, 1956, pp. 62–66.

(7) R. E. Moore, Interval Arithmetic and Automatic Error Analysis in Digital Computing, Technical Report No. 25, Stanford Univ., Stanford, California, 1962, p. 9.

(8) Ibid., p. 10.

(9) Ibid., pp. 16–17.

(10) Ibid., pp. 21–39.

(11) Ibid., pp. 58–71.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abramowitz, M. and Stegun, I. A. (ed.), Handbook of Mathematical Functions, AMS 55, National Bureau of Standards, Washington, D. C., 1964.

Aliev, R. G., "On the Question of Specific Criteria for Estimates of the Lengths of Subcritical Intervals," Izv. VUZ, Matematika, No. 5 (42), 1964, pp. 3–7.

Aliev, R. G., Ostroumov, V. V., and Pak, S. A., "On Certain Properties of Cauchy Functions," Izv. VUZ, Matematika, No. 4 (41), 1964, pp. 9–11.

Birkhoff, G. and MacLane, S., A Survey of Modern Algebra, Revised Edition, New York, 1953.

Coddington, E. A., An Introduction to Ordinary Differential Equations, New York, 1961.

Collins, George, Interval Arithmetic for Automatic Error Analysis, International Business Machines Corp., Mathematics and Applications Dept., New York, 1960.

Dwyer, P. S., Linear Computations, New York, 1951.

Eisenhart, L. P., Continuous Groups of Transformations, Princeton, New Jersey, 1933.

Gendzhoian, G. V., "On the Two-Sided Chaplygin Approximation of the Solution of Two-Point Boundary Problems," Izv. AN Arm SSR, Fiz. Matematika Nauk, Vol. 17, No. 3, 1964, pp. 21–26.

Gorn, Saul, "The Automatic Analysis and Control of Computing Errors," J. Soc. Indus. Appl. Math., Vol. 2, No. 2, June 1954, pp. 69–81.

Hildebrand, F. B., Introduction to Numerical Analysis, New York, 1956.

Hille, E. and Phillips, R. S., Functional Analysis and Semi-Groups, Providence, Rhode Island, 1957.

Ince, E. L., Ordinary Differential Equations, New York, 1956.

Kaplan, Wilfred, Advanced Calculus, Cambridge, Massachusetts, 1953.

Kaplan, Wilfred, Ordinary Differential Equations, Reading, Massachusetts, 1958.

Kelley, J. L., General Topology, New Jersey, 1955.

Kolmogorov, A. N. and Fomin, S. V., Elements of the Theory of Functions and Functional Analysis, Vol. 1, Translated by L. F. Boron, Rochester, New York, 1957. (Original Russian publication 1954.)

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

Moore, R. E., Interval Arithmetic and Automatic Error Analysis in Digital Computing, Technical Report No. 25, Applied Mathematics and Statistics Laboratories, Stanford University, Stanford, California, 1962.

Moore, R. E., Strother, W., and Yang, C. T., Interval Analysis, LMSD-703073, Lockheed Missiles & Space Company, Sunnyvale, California, 1960.

Moore, R. E. and Yang, C. T., Interval Analysis I, LMSD-285875, Lockheed Missiles & Space Company, Sunnyvale, California, 1959.

Murray, F. J. and Miller, K. S., Existence Theorems for Ordinary Differential Equations, New York, 1954.

Nelson, A. L., Folley, K. W. and Coral, M., Differential Equations, Second Edition, Boston, 1960.

Pak, S. A. and Chichkin, E. S., "On the Existence of Upper and Lower Solutions of the Cauchy Problem of a Second Order Differential Equation," Izv. VUZ, Matematika, No. 5 (42), 1964, pp. 91–94.

Taylor, A. E., Advanced Calculus, Boston, 1955.

APPENDIX A

APPENDIX A

Round-Off Error In Interval Arithmetic

A complete listing of an interval arithmetic computer program for the IBM 7094 is given in the next section. The program is written in FAP, the IBM 7094 machine language. It can be used as a subroutine with a FORTRAN main program and reached via a CALL to the appropriate operation.

Of particular interest are the sections which do the rounding after an operation has been completed. To insure that the closed interval contains the required result, round-off error must be carefully considered.

For example, the interval containing the square root of the closed interval $[1, 2]$ must contain all values of x for $\sqrt{1} \leq x \leq \sqrt{2}$. But $\sqrt{2}$ is an irrational number and cannot be represented exactly in any digital computer. Therefore, when we find the number which is approximately equal to the square root of two, we must be sure that if the number is squared we obtain at least two. Otherwise we must add enough to have the square equal at least two. Clearly, if the square of the number approximating $\sqrt{2}$ were slightly less than $\sqrt{2}$, then the exact value would not be included within the interval. This, of course, is the case where $\sqrt{2}$ is the right end of the interval.

If, however, the interval was $[\sqrt{2}, \sqrt{4}]$, we would want the number approximating $\sqrt{2}$ to be just less than two when squared, so as to be sure that all values of x for $\sqrt{2} \leq x \leq \sqrt{4}$ where included in the interval.

The same round-off problem occurs with all interval arithmetic operations and the particular digital computer being used is of no great consequence, since all computers have finite word lengths.

An Interval Arithmetic Computer Program

The following computer program for the IBM 7094 will add, subtract, multiply, and divide two closed intervals, take the square root of a closed interval and give the inverse of a closed interval. A similar program for any other digital computer can also be programmed.

```
*       FAP
*
*       INTERVAL ARITHMETIC PROGRAM
*
        ENTRY   ADD
        ENTRY   SUB
        ENTRY   MULT
        ENTRY   DIV
        ENTRY   INVS
        ENTRY   SQR
        ENTRY   ROVFL
        ENTRY   RCEL8
*
*
*       THE FOLLOWING HOLDS FOR ADD, SUB, MULT, AND DIV
*       A IN ACC, B IN ACC+1
*       C IN OPER, D IN OPER+1
*       RESULT IN ACC AND ACC+1
*
*       INTERVAL SUBTRACTION  (A,B)-(C,D)=(A,B)+(-D,-C)
*
SUB     CLS  OPER
        XCA
        CLS  OPER+1              NOTE - OPER AND OPER+1 HAVE
        STO  OPER   -D                  BEEN MODIFIED
        STQ  OPER+1 -C
*
*       INTERVAL ADDITION  (A,B)+(C,D)=(A+C,B+D)
*
ADD     CLA  ACC
        STO  TEMP   FOR ROUNDING
        FAD  OPER   ADD LEFT ENDS
        STO  ACC    LEFT END SUM
        TPL  *+2    IF LEFT END IS POSITIVE, NO ROUNDING NEEDED
        TSX  RLDA,2 LEFT END NEGATIVE, ROUND DOWN
        CLA  ACC+1  ADD RIGHT ENDS
        STO  TEMP   FOR ROUNDING
        FAD  OPER+1
```

```
        STO   ACC+1   RIGHT END SUM
        TMI   1,4     RETURN TO MAIN PROGRAM
        TRA   RRUA    RIGHT END POSITIVE, ROUND UP
*
*       INTERVAL DIVISION  (A,B)/(C,D)=(A,B)(1/D,1/C)
*                          C NOT=0, D NOT=0
*                          C,D MUST HAVE SAME SIGN
*       THE DIV ROUTINE WILL GIVE THE INVERSE OF (C,D)
*         WE THEN GO INTO THE MULT ROUTINE
*
*   NOTE - OPER AND OPER+1 HAVE BEEN MODIFIED.
*
DIV     CLA   OPER+1
        TZE   ERR     D=0
        XCA           D NOT=0
        CLA   OPER
        TZE   ERR1    C=0
        TQP   *+3     C NOT=0
        TMI   *+3     D-,C-   OK
        TRA   ERR3    D-,C+   ILLEGAL INTERVAL
        TMI   ERR+2   D+,C-  INTERVAL INCLUDES ZERO
        CLA   FLPT1   C,D HAVE SAME SIGN, NEITHER = 0
        FDH   OPER+1  (1/D)
        STQ   OPER+1  TEMPORARY STORE OF NEW LEFT END, UNROUNDED
        TQP   *+9     QUOTIENT PLUS, NO ROUNDING NEEDED
        XCA           QUOTIENT NEGATIVE, ROUND
        LLS   8       SHIFT OFF CHARACTERISTIC OF REMAINDER
        XCA           SIGN + FRACTION OF REMAINDER TO AC
        TZE   *+5     REMAINDER IS ZERO, NO ROUNDING NEEDED
        CLA   OPER+1  QUOTIENT TO BE ROUNDED DOWN, KNOWN NEGATIVE
        LDQ   MASK1   PUT 1 BIT IN POSITION 9 FOR ROUNDING
        FRN
        STO   OPER+1  (1/D) - ROUNDED
        CLA   FLPT1   GET NEW RIGHT END
        FDH   OPER    (1/C)
        STQ   OPER
        XCA           QUOTIENT TO AC
        TMI   *+8     QUOTIENT NEGATIVE, NO ROUNDING NEEDED
        LLS   8       SHIFT OFF CHARACTERISTIC OF REMAINDER
        XCA           SIGN + FRACTION OF REMAINDER TO AC
        TZE   *+5     REMAINDER IS ZERO, NO ROUNDING NEEDED
        CLA   OPER    QUOTIENT TO BE ROUNDED UP, KNOWN POSITIVE
        LDQ   MASK1   PUT 1 BIT IN POSITION 9 FOR ROUNDING
        FRN
        TRA   *+2     (1/C) - ROUNDED IN AC
        CLA   OPER
        LDQ   OPER+1  (1/D)
        STQ   OPER    (1/D) - NEW LEFT END
        STO   OPER+1  (1/C) - NEW RIGHT END
*
*       INTERVAL MULTIPLICATION
```

```
*          (A,B)(C,D)=(MIN(AC,AD,BC,BD),MAX(AC,AD,BC,BD))
*   NOTE -- THE MULTIPLICATION ROUTINE USES THE INDICATORS.
*
MULT    CLA   ACC     FETCH A                          A IN ACC
        TMI   MULT3   A-                               B IN ACC+1
        LDQ   OPER    A+,B+    TEST C                  C IN OPER
        TQP   MULT2   C+                               D IN OPER+1
        LDQ   OPER+1  A+,B+,C-   TEST D
        TQP   MULT1   A+,B+,C-,D+
        FMP   ACC     A+,B+,C-,D-
        STO   TEMP    (AD)
        LDQ   ACC+1   B+
        FMP   OPER    C-
        STO   ACC     (BC)
        TSX   RLD,2   ROUND LEFT END DOWN
        CLA   TEMP
        STO   ACC+1   (AD)
        TRA   1,4     EXIT  (BC,AD)  A+,B+,C-,D-
MULT1   LDQ   ACC+1   A+,B+,C-,D+
        FMP   OPER
        STO   ACC     (BC)
        TSX   RLD,2   ROUND LEFT END DOWN
        LDQ   ACC+1
        FMP   OPER+1
        STO   ACC+1
        TRA   RRU     ROUND RIGHT END UP AND EXIT, (BC,BD)
MULT2   FMP   ACC     A+,B+,C+,D+
        STO   ACC     (AC)
        LDQ   OPER+1
        FMP   ACC+1
        STO   ACC+1   (BD)
        TRA   RRU     ROUND RIGHT END UP AND EXIT, (AC,BD)
MULT3   CLA   ACC+1   A-, TEST B
        TMI   MULT6   B-
        LDQ   OPER    A-, B+, TEST C
        TQP   MULT5   A-, B+, C+
        LDQ   OPER+1  A-, B+, C-, TEST D
        TQP   MULT4   A-, B+, C-, D+
        CLA   ACC     A-, B+, C-, D-,  GET (BC,AC)
        STO   TEMP
        LDQ   ACC+1
        FMP   OPER
        STO   ACC     (BC)
        TSX   RLD,2   ROUND LEFT END DOWN
        LDQ   TEMP
        FMP   OPER
        STO   ACC+1   (AC)
        TRA   RRU     ROUND RIGHT END UP AND EXIT, (BC,AC)
MULT4   FMP   ACC     A-, B+, C-, D+, INDETERMINABLE BY SIGN TEST
        STO   TEMP    (AD)
        STQ   TEMP+1
```

```
        LDQ   ACC+1
        FMP   OPER    (BC)
        CAS   TEMP
        TRA   *+3     AC GREATER THAN TEMP, STORE (AD)
        TRA   *+5     AC EQUAL TO TEMP
        TRA   *+13    AC LESS THAN TEMP, STORE (BC)
        CLA   TEMP
        LDQ   TEMP+1
        TRA   *+10
        XCA           MOST SIGNIF PART TO MQ, LEAST SIGNIF TO AC
        CAS   TEMP+1  MOST SIGNIF PARTS EQUAL, TEST LEAST SIGNIF
        TRA   *+3     AC GREATER THAN TEMP+1, STORE (AD)
        TRA   *+5     AC EQUAL TO TEMP+1, STORE (BC)
        TRA   *+4     AC LESS THAN TEMP+1, STORE (BC)
        CLA   TEMP
        LDQ   TEMP+1  REPLACE AC AND MQ BY TEMP AND TEMP+1
        TRA   *+2
        XCA           MOST SIGNIF PART TO AC, LEAST SIGNIF TO MQ
        LDI   ACC     PUT ACC (A) IN INDICATORS FOR TEMP STORAGE
        STO   ACC     THE SMALLER OF (AD) AND (BC)
        TSX   RLD,2   ROUND LEFT END DOWN
        PIA           PUT A IN AC, FIND RIGHT END
        XCL
        FMP   OPER
        STO   TEMP    (AC)
        STQ   TEMP+1
        LDQ   ACC+1
        FMP   OPER+1  (BD)
        CAS   TEMP
        TRA   *+13    AC GREATER THAN TEMP, STORE (BD)
        TRA   *+4     AC EQUAL TEMP
        CLA   TEMP    AC LESS THAN TEMP, STORE (AC)
        LDQ   TEMP+1
        TRA   *+9
        XCA           MOST SIGNIF PART TO MQ, LEAST SIGNIF TO AC
        CAS   TEMP+1  MOST SIGNIF PARTS EQUAL, TEST LEAST SIGNIF
        TRA   *+5     AC GREATER THAN TEMP+1, STORE (BD)
        TRA   *+4     AC EQUAL TO TEMP+1, STORE (BD)
        CLA   TEMP    AC LESS THAN TEMP+1, STORE (AC)
        LDQ   TEMP+1  REPLACE AC AND MQ BY TEMP AND TEMP+1
        TRA   *+2
        XCA           MOST SIGNIF PART TO AC, LEAST SIGNIF TO MQ
        STO   ACC+1   THE LARGER OF (AC) AND (BD)
        TRA   RRU     ROUND RIGHT END UP AND EXIT
MULT5   LDQ   ACC     A-,B+,C+,D+
        FMP   OPER+1
        STO   ACC     (AD)
        TSX   RLD,2   ROUND LEFT END DOWN
        LDQ   ACC+1
        FMP   OPER+1
        STO   ACC+1   (BD)
```

```
         TRA   RRU     ROUND RIGHT END UP AND EXIT,  (AD,BD)
MULT6    CLA   OPER    A-,B-, TEST C
         TMI   MULT7   A-,B-,C-
         LDQ   ACC     A-,B-,C+,D+   GET (AD,BC)
         FMP   OPER+1
         STO   ACC     (AD)
         TSX   RLD,2   ROUND LEFT END DOWN
         LDQ   ACC+1
         FMP   OPER
         STO   ACC+1   (BC)
         TRA   1,4     EXIT          (AD,BC)
MULT7    CLA   ACC     ACC STORED IN TEMP WHETHER D IS + OR -
         STO   TEMP
         LDQ   OPER+1  A-,B-,C-, TEST D
         TQP   MULT8   A-,B-,C-,D+
         FMP   ACC+1   A-,B-,C-,D-   GET (BD,AC)
         STO   ACC     (BD), NO ROUNDING NEEDED
         LDQ   TEMP
         FMP   OPER
         STO   ACC+1   (AC)
         TRA   RRU     ROUND RIGHT END UP AND EXIT, (BD,AC)
MULT8    FMP   ACC     A-,B-,C-,D+ ,  GET (AD,AC)
         STO   ACC     (AD)
         TSX   RLD,2   ROUND LEFT END DOWN
         LDQ   TEMP
         FMP   OPER
         STO   ACC+1   (AC)
         TRA   RRU     ROUND RIGHT END UP END EXIT, (AD,AC)
*
*        INTERVAL INVERSE  (1,1)/(C,D)=(1/D,1/C)
*                           C NOT=0, D NOT=0
*                           C,D MUST HAVE SAME SIGN
*        C IN OPER, D IN OPER+1
*        RESULT IN ACC AND ACC+1
*
INVS     CLA   OPER+1
         TZE   ERR     D=0
         XCA           D NOT=0
         CLA   OPER
         TZE   ERR1    C=0
         TQP   *+3     C NOT=0
         TMI   *+3     D-,C-   OK
         TRA   ERR3    D-,C+   ILLEGAL INTERVAL
         TMI   ERR2    D+,C-  LEGAL INTERVAL, BUT NOT FOR DIVISION
         CLA   FLPT1   C,D HAVE SAME SIGN, NEITHER = 0
         FDH   OPER+1  (1/D)
         STQ   OPER+1  TEMPORARY STORE OF NEW LEFT END, UNROUNDED
         TQP   *+9     QUOTIENT PLUS, NO ROUNDING NEEDED
         XCA           QUOTIENT NEGATIVE, ROUND
         LLS   8       SHIFT OFF CHARACTERISTIC OF REMAINDER
         XCA           SIGN + FRACTION OF REMAINDER TO AC
```

```
        TZE   *+5     REMAINDER IS ZERO, NO ROUNDING NEEDED
        CLA   OPER+1  QUOTIENT TO BE ROUNDED DOWN, KNOWN NEGATIVE
        LDQ   MASK1   PUT 1 BIT IN POSITION 9 FOR ROUNDING
        FRN
        STO   OPER+1  (1/D) - ROUNDED
        CLA   FLPT1   GET NEW RIGHT END
        FDH   OPER    (1/C)
        STQ   OPER
        XCA           QUOTIENT TO AC
        TMI   *+8     QUOTIENT NEGATIVE, NO ROUNDING NEEDED
        LLS   8       SHIFT OFF CHARACTERISTIC OF REMAINDER
        XCA           SIGN + FRACTION OF REMAINDER TO AC
        TZE   *+5     REMAINDER IS ZERO, NO ROUNDING NEEDED
        CLA   OPER    QUOTIENT TO BE ROUNDED UP, KNOWN POSITIVE
        LDQ   MASK1   PUT 1 BIT IN POSITION 9 FOR ROUNDING
        FRN
        TRA   *+2     (1/C) - ROUNDED IN AC
        CLA   OPER
        LDQ   OPER+1  (1/D)
        STQ   ACC     (1/D)
        STO   ACC+1   (1/C)
        TRA   1,4     RETURN TO MAIN PROGRAM
*
*       INTERVAL SQUARE ROOT, SQUARE ROOT OF (A,B)
*         A IN OPER, B IN OPER+1
*         SQ RT OF A IN ACC, SQ RT OF B IN ACC+1
*         LEFT END, IF SQ OF SQ RT -
*           IS GREATER THAN A, SUB 1 BIT FROM POSITION 35 AND
*             COMPARE AGAIN
*           IS EQUAL TO A, CHECK LEAST SIGNIF PORTION, IF LSP
*             =0 ACCEPT SQ RT, IF LSP NOT=0, SUB 1 BIT FROM
*             POSITION 35 AND ACCEPT AS SQ RT
*           IS LESS THAN A, ACCEPT SQ RT (SQ RT IS LESS THAN
*             1 BIT FROM EXACT VALUE)
*         RIGHT END, IF SQ RT -
*           IS GREATER THAN B, SUB 1 BIT FROM POSITION 35 AND
*             COMPARE AGAIN
*           IS EQUAL TO B, ACCEPT AS SQ RT
*           IS LESS THAN B, ADD 1 BIT TO POSITION 35 AND
*             ACCEPT AS SQ RT
*
SQR     CLA   OPER    LEFT END, FETCH A
        TNZ   *+3
        STZ   ACC     SQ RT OF ZERO IS ZERO
        TRA   SQRT3   GO WORK ON RIGHT END
        TMI   ERR5    NEGATIVE NUMBER
        TSX   SRT,2   GET SQ RT OF LEFT END
SQRT1   STO   TEMP    SQ RT OF A
        XCA
        FMP   TEMP    SQ OF SQ RT
        CAS   OPER    COMPARE WITH A
```

```
        TRA   SQRT2  SQ OF SQ RT IS GREATER THAN A
        TRA   *+4                 IS EQUAL TO A
        CLA   TEMP                IS LESS THAN A, STORE SQRT
        STO   ACC    STORE LEFT END
        TRA   SQRT3  GO ON TO RIGHT END
        LGL   9      CHECK LEAST SIGNIF PORTION, SHIFT OF CHAR
        XCL
        TZE   *-5    LEAST SIGNIF POTION = 0, STORE SQRT
        CLA   TEMP   PICK UP SQRT
        ANA   MASK   KEEP CHARACTERISTIC, MASK OFF FRACTION
        ADD   ONE    1 BIT IN POSITION 35
        SSM
        FAD   TEMP   SUB 1 BIT FROM POSITION 35
        STO   ACC    STORE LEFT END
        TRA   SQRT3  GO ON TO RIGHT END
SQRT2   CLA   TEMP   SUB BIT FROM POSITION 35 AND COMPARE AGAIN
        ANA   MASK   KEEP CHAR, MASK OFF FRACTION
        ADD   ONE    1 BIT IN POSITION 35
        SSM
        FAD   TEMP   SUB 1 BIT FROM POSITION 35
        TRA   SQRT1  GO SQ, THEN COMPARE AGAIN
SQRT3   CLA   OPER+1 RIGHT END, FETCH B
        TNZ   *+3
        STZ   ACC+1  SQRT OF ZERO IS ZERO
        TRA   1,4    EXIT
        TSX   SRT,2  GET SQ RT OF RIGHT END
SQRT4   STO   TEMP   SQRT OF B
        XCA
        FMP   TEMP   SQ OF SQRT
        CAS   OPER+1 COMPARE WITH B
        TRA   SQRT5  SQ OF SQRT IS GREATER THAN B
        TRA   *+6                 IS EQUAL TO B, STORE SQRT
        CLA   TEMP                IS LESS THAN B, ADD 1 BIT TO
        LDQ   MASK1  1 BIT IN 9 OF MQ      POSITION 35 AND STORE
        FRN          ROUND - ADD A BIT
        STO   ACC+1  STORE RIGHT END
        TRA   1,4    EXIT
        CLA   TEMP   SQRT OF B
        STO   ACC+1  STORE RIGHT END
        TRA   1,4    EXIT
SQRT5   CLA   TEMP   SUB BIT FROM POSITION 35 AND COMPARE AGAIN
        ANA   MASK   KEEP CHAR, MASK OFF FRACTION
        ADD   ONE    1 BIT IN POSITION 35
        SSM
        FAD   TEMP   SUB 1 BIT FROM POSITION 35
        TRA   SQRT4  GO SQ, THEN COMPARE AGAIN
*
*       ENTER WITH ARGUEMENT IN AC, RETURN WITH SQ RT IN AC
*
SRT     SXA   *+16,2
        AXT   3,2
```

```
        STO   TEMP    X
        ANA   *+16    LAST DIGIT OF POWER
        ARS   1
        ADD   TEMP    X
        ARS   1
        ADD   *+11    100.4
        STO   TEMP+1  Y1, ETC
        CLA   TEMP    X
        FDH   TEMP+1  X/Y1 = Q1
        CLA   TEMP+1  Y1
        STQ   TEMP+1  Q1
        FAD   TEMP+1  Y1+Q1
        SUB   *+5     DIVIDE BY 2
        TIX   *-7,2
        AXT   **,2
        TRA   1,2
        OCT   100400000000
        OCT   001000000000
*
*       THIS SUBROUTINE IS USED ONLY FOR ROUNDING ADD AND SUB
*       ROUND LEFT END DOWN (ONLY WHEN LEFT END IS KNOWN TO
*         BE NEGATIVE).  IF REMAINDER IS ZERO, DO NOT ROUND.
*         REMAINDER IS IN MQ, MOST SIGNIFICANT PART IS IN
*         ACC.  RETURN IS BY TRA 1,2 TO WORK ON RIGHT END
*
RLDA    LLS   8       SHIFT OFF CHARACTERISTIC OF REMAINDER
        XCA           SIGN + FRACTION OF REMAINDER TO AC
        TNZ   RLDA1   REMAINDER NON-ZERO, ROUND
        NZT   OPER    REMAINDER ZERO, CHECK FURTHER, IS OPER ZERO
        TRA   1,2      YES, NO ROUNDING NEEDED, RETURN TO ADD
        NZT   TEMP     NO, IS TEMP ZERO (TEMP HAS ORIGINAL ACC)
        TRA   1,2      YES, NO ROUNDING NEEDED, RETURN TO ADD
        CAL   TEMP     NO, IS DIFF OF CHARACTERISTICS GREATER
        ANA   MASK    MASK OFF SIGN AND FRACTION      THAN (53)10
        STO   TEMP+1   KEEP CHARACTERISTIC
        CAL   OPER    MASK OFF SIGN AND FRACTION
        ANA   MASK     KEEP CHARACTERISTIC
        SUB   TEMP+1  DIFFERENCE OF THE CHARACTERISTICS
        LAS   OCT66   COMPARE ABSOLUTE VALUE WITH (66)8
        TRA   *+3      GREATER THAN (66)8, ROUND
        TRA   *+2      EQUAL TO (66)8, ROUND
        TRA   1,2      LESS THAN (66)8, NO ROUNDING RETURN TO ADD
RLDA1   CLA   ACC     QUANTITY TO BE ROUNDED DOWN, KNOWN NEG
        LDQ   MASK1   PUT 1 BIT IN POSITION 9 FOR ROUNDING
        FRN
        STO   ACC     ROUNDED LEFT END
        TRA   1,2     RETURN TO ADD
*
*       THIS SUBROUTINE IS USED ONLY FOR ROUNDING ADD AND SUB
*       ROUND RIGHT END UP (ONLY WHEN RIGHT END IS KNOWN TO
*         BE POSITIVE).  IF REMAINDER IS ZERO, DO NOT ROUND.
```

```
*         REMAINDER IS IN MQ, MOST SIGNIFICANT PART IS IN
*         ACC+1.  RETURN IS BY TRA 1,4 TO MAIN PROGRAM.
*
RRUA    LLS   8       SHIFT OFF CHARACTERISTIC OF REMAINDER
        XCA           SIGN + FRACTION OF REMAINDER TO AC
        TNZ   RRUA1   REMAINDER NON-ZERO, ROUND
        NZT   OPER+1  REMAINDER ZERO, CHECK FURTHER, IS OPER ZERO
        TRA   1,4      YES, NO ROUNDING NEEDED, EXIT
        NZT   TEMP     NO, IS TEMP ZERO (TEMP HAS ORIGINAL ACC)
        TRA   1,4      YES, NO ROUNDING NEEDED, EXIT
        CAL   TEMP     NO, IS DIFF OF CHARACTERISTICS GREATER
        ANA   MASK    MASK OFF SIGN AND FRACTION      THAN (53)10
        STO   TEMP+1   KEEP CHARACTERISTIC
        CAL   OPER+1  MASK OFF SIGN AND FRACTION
        ANA   MASK     KEEP CHARACTERISTIC
        SUB   TEMP+1  DIFFERENCE OF THE CHARACTERISTICS
        LAS   OCT66   COMPARE ABSOLUTE VALUE WITH (66)8
        TRA   *+3      GREATER THAN (66)8, ROUND
        TRA   *+2      EQUAL TO (66)8, ROUND
        TRA   1,4      LESS THAN (66)8, NO ROUNDING, EXIT
RRUA1   CLA   ACC+1   QUANTITY TO BE ROUNDED UP, KNOWN POS
        LDQ   MASK1   PUT 1 BIT IN POSITION 9 FOR ROUNDING
        FRN
        STO   ACC+1   ROUNDED RIGHT END
        TRA   1,4     RETURN TO MAIN PROGRAM
*
*       THIS SUBROUTINE IS NOT USED BY ADD OR SUB
*       ROUND LEFT END DOWN (ONLY WHEN LEFT END IS KNOWN TO
*         BE NEGATIVE).  IF REMAINDER IS ZERO, DO NOT ROUND.
*         REMAINDER IS IN MQ, MOST SIGNIFICANT PART IS IN
*         ACC.  RETURN IS BY TRA 1,2 TO WORK ON RIGHT END
*
RLD     LLS   8       SHIFT OFF CHARACTERISTIC OF REMAINDER
        XCA           SIGN + FRACTION OF REMAINDER TO AC
        TZE   1,2     REMAINDER IS ZERO, GO WORK ON RIGHT ENDS
        CLA   ACC     QUANTITY TO BE ROUNDED DOWN - KNOWN NEG
        LDQ   MASK1   PUT 1 BIT IN POSITION 9 FOR ROUNDING
        FRN
        STO   ACC     ROUNDED LEFT END
        TRA   1,2
*
*       THIS SUBROUTINE IS NOT USED BY ADD OR SUB
*       ROUND RIGHT END UP (ONLY WHEN RIGHT END IS KNOWN TO
*         BE POSITIVE).  IF REMAINDER IS ZERO, DO NOT ROUND.
*         REMAINDER IS IN MQ, MOST SIGNIFICANT PART IS IN
*         ACC+1.  RETURN IS BY TRA 1,4 TO MAIN PROGRAM.
*
RRU     LLS   8       SHIFT OFF CHARACTERISTIC OF REMAINDER
        XCA           SIGN + FRACTION OF REMAINDER TO AC
        TZE   1,4     REMAINDER IS ZERO, EXIT TO MAIN PROGRAM
        CLA   ACC+1   QUANTITY TO BE ROUNDED UP - KNOWN POSITIVE
```

```
        LDQ   MASK1   PUT 1 BIT IN POSITION 9 FOR ROUNDING
        FRN
        STO   ACC+1   ROUNDED RIGHT END
        TRA   1,4
*       MAIN PROGRAM SHOULD HAVE A   CALL ROVFL   WHICH WILL
*         SET CELL 8 TO HANDLE FLOATING POINT SPILLS DURING
*         EXECUTION OF INTERVAL ARITHMETIC.  AT CONCLUSION OF
*         JOB, THE MONITOR WILL RESET CELL 8 FOR THE NORMAL
*         HANDLING OF SPILLS, SO THAT IT IS NOT NECESSARY TO
*         RESTORE CELL 8.  HOWEVER SHOULD IT BE DESIRED THAT
*         CELL 8 BE RESTORED FOR NORMAL HANDLING OF FLOATING
*         POINT SPILLS OF NON-INTERVAL ARITHMETIC CALCULA-
*         TIONS, A   CALL RCEL8   WILL RESTORE CELL 8 AND
*         THUS RESTORE THE NORMAL HANDLING OF SPILLS.
*
ROVFL   CLA   8       SAVE
        STO   *+5           CELL 8
        CLA   *+3
        STO   8
        TRA   1,4     EXIT
        TRA   ERR4    TRA TO BE PLACED IN CELL 8
        PZE           CELL 8 SAVED HERE
RCEL8   CLA   *-1     RESET CELL 8
        STO   8
        TRA   1,4     EXIT
*
ERR     HTR   *       D=0     DIVIDE ERROR
ERR1    HTR   *       D=0     DIVIDE ERROR
ERR2    HTR   *       C-,D+   DIVIDE ERROR, INTERVAL CONTAINS 0
ERR3    HTR   *       C+,D-   DIVIDE ERROR, ILLEGAL INTERVAL
ERR4    HTR   *       FLOATING POINT SPILL
ERR5    HTR   *       SQRT OF NEGATIVE ARGUMENT
*
*
FLPT1   DEC   1,0
MASK    OCT   377000000000
MASK1   OCT   000400000000        1 BIT IN POSITION 9
OCT66   OCT   066000000000        USED IN ROUNDING ADD AND SUB
ONE     OCT   1
TEMP    BSS   2
*
        COMMON  1
ACC     COMMON  2
OPER    COMMON  1
        END
```