This file contains:

- a copy of the title page from the book in which the subsequent paper appears; and
- a copy of the paper, itself.

Error in Digital Computation
Volume 1

Proceedings of an Advanced Seminar Conducted
by the Mathematics Research Center,
United States Army, at the University
of Wisconsin, Madison, October 5-7, 1964.

Edited by

Louis B. Rall

JOHN WILEY & SONS, INC.
New York • London • Sydney

Printed in the United States of America

RAMON E. MOORE

The Automatic Analysis and Control of Error in Digital Computation Based on the Use of Interval Numbers

1. Introduction

A digital computation is a finite sequence of inexact arithmetic operations.

In order to approximate quantities defined by finite or infinite sequences of exact arithmetic operations, approximation schemes are designed in the form of digital computations containing a number of "approximation parameters" such as: the number of digits to be carried in limited precision arithmetic, the number of terms to be carried in a truncated infinite series, the "step size" to be used in a discretization approximation to an integration, the number of iterations in an iterative process, etc.

Each choice of a set of values for the approximation parameters will give a definite digital computation leading to results with certain errors.

The analysis and control of error in digital computation involves choosing a set of parameter values which lead to results with tolerable errors. There are usually infinitely many sets of such choices and we further require that the total elapse of time during the execution of the computation be about as small as possible. Of course, there are usually constraints on the choice of parameter values such as limited storage space in the computer.

If we can mechanize the determination of appropriate approximation parameter values, then we can take fuller advantage of the great speed of automatic digital computers [13], [14], [20], [23]. Our approach to a solution of this problem is to start with the things we know a computer can do.

In section 2, a brief examination of the nature of digital computation leads to the concept of "interval numbers". This is essentially the same concept that is used in experimental scientific work in representing a quantity known only within a certain accuracy in the form $x \pm \epsilon$.

In sections 2 and 3, arithmetic operations with interval numbers are discussed and it is seen that "rounded" interval arithmetic is a means for the automatic determination of upper bounds to the accumulated round-off error in any digital computation.

In section 4, a convergence theorem is proved providing for the computation of intervals containing and arbitrarily close to the exact range of values of real rational functions on finite intervals.

In section 5, some iterative computations with intervals are shown to produce sequences of intervals contracting to roots of algebraic equations.

In section 6, continuity is defined for interval functions and a geometric picture of the set of interval numbers is given.

In section 7, a convergence theorem is proved providing for the computation of intervals of arbitrarily small width containing the exact value of a definite integral. A more general theorem is then proved in which the order of convergence can be chosen arbitrarily. An interval version of Gaussian quadrature is also presented.

Our main result, presented in sections 8-11, is the application of interval computation to the design of a computer program [24] for the initial value problem in ordinary differential equations. In carrying through this application in detail, some additional new techniques are presented including a practical procedure for the recursive generation of Taylor coefficients.

The resulting program automatically determines all required approximation parameter values except "word length".

Given only a system of differential equations and initial conditions, the program produces intervals containing exact solution values at any value of the independent variable within a certain distance of the nearest singular point or until, for whatever reason, numbers occur in the computation exceeding the range of machine representable numbers for a given word length. Numerical results are given illustrating "single" and "double" precision floating point versions which have been coded.

2. Interval Numbers

The first thing we want is a means by which the computer can say how far off the results of its limited precision arithmetic operations can be. The answer to such a question can have two forms each consisting of a pair of machine numbers: (1) an approximate result and an error bound, (2) a lower and an upper bound to the exact result.

In either case, the answer has the form of an interval which contains the exact result, $[x-\epsilon, x+\epsilon]$ or $[a, b]$ where x, ϵ, a, b are numbers computed by the machine.

In order to enable the computer to bound the accumulation of error, if we wish, say, the computer to determine the sum of two numbers y_1, y_2 where it is only known that $y_1 \in [a_1, b_1]$ and $y_2 \in [a_2, b_2]$, then the best we could do would be to have the computer determine that

$$y_1 + y_2 \in [a_1 + a_2,\ b_1 + b_2] .$$

Actually, since the computer cannot even determine $a_1 + a_2$, or $b_1 + b_2$ exactly, in general, we may have to further adjust the last digit of one or both end points of the machine values for $a_1 + a_2, b_1 + b_2$. The details of how this is done will depend on the arithmetic operations in a particular computer. If an unrounded single precision result is computed, for example, with a true remainder, of the same sign as the result, available for testing, then we would round positive right ends and negative left ends of intervals away from zero, but not round positive left ends or negative right ends or any end points for which a zero remainder was obtained.

In any case, with the rounding procedures dependent on the detail of the arithmetic used on a particular computer, machine programs can be written which produce an interval K containing the results of an exact arithmetic operation on any pair of numbers one from an interval I and one from an interval J, the end points of all these intervals being numbers representable on the computer. A number of such programs have, in fact, been written, [1], [4], [10], [12], [20],...

Now, each finite sequence of arithmetic operations in a computer program defines a rational function, whose values are computed by executing that part of the program.

We wish to enable the computer to bound the range of values of any rational function over given ranges of values for its arguments. In particular, this will enable us to bound programmed expressions for remainder terms in the truncation of infinite processes. A remainder term may be an expression involving a variable ξ, for example, which is only known to lie within a certain invertal $a \leq \xi \leq b$. Therefore, even leaving rounding errors aside, we still wish to be able to bound the results of arithmetic operations on quantities known only to lie within certain intervals.

Thus, the actual processes of digital computation lead in a natural way to the consideration of an arithmetic system which operates with intervals of real numbers. The idea of computation with interval numbers (or "range" numbers), especially in connection with computers, has been entertained recently by a number of authors, [1],[4], [6], [7],[8],[10],[12],[20],[23],[24],[32],...

Of course, in a more general setting, mathematicians have for some time been dealing with set-valued functions and commonly use

such notation as $f(A)$ to represent the set of values of $f(x)$ for $x \in A$, or the "image of A under f". Interval valued functions and functions of intervals in particular have been considered, [2], [3], [21], [22], [23], [27], [28]. Set-valued algebras have also been considered, [33].

One of the essential features of intervals in connection with computing is that they are represented by finite sets of numbers, in fact by pairs of numbers.

3. Interval Arithmetic

In this section we will define an arithmetic system operating with intervals of real numbers and we will study some of the properties of that arithmetic.

If $*$ is one of the symbols $+, -, \cdot, \div$, and $[a,b]$, $[c,d]$ are closed intervals of real numbers, then the arithmetic operations on interval numbers are defined by

$$[a,b] * [c,d] = \{x*y \mid a \leq x \leq b,\ c \leq y \leq d\}, \tag{3.1}$$

except that we define

$$[a,b] \div [c,d]$$

only in case $0 \notin [c,d]$.

Since the real arithmetic operations are continuous, they map the compact connected sets $[a,b] \otimes [c,d]$, ($\otimes$ denotes the Cartesian product), onto compact connected sets, i.e., closed real intervals. In fact, we have the formulas

$$\begin{aligned} [a,b] + [c,d] &= [a+c, b+d] \\ [a,b] - [c,d] &= [a-d, b-c] \\ [a,b] \cdot [c,d] &= [\min(ac, ad, bc, bd), \max(ac, ad, bc, bd)] \end{aligned} \tag{3.2}$$

and if $0 \notin [c,d]$, then

$$[a,b] \div [c,d] = [a,b] \cdot [1/d, 1/c].$$

In the case of interval multiplication, by examining the signs of a, b, c, d only two multiplications need be carried out to determine $[a,b] \cdot [c,d]$ except in the case $a < 0 < b$, $c < 0 < d$ where $[a,b] \cdot [c,d] = [\min(ad, bc), \max(ac, bd)]$.

It follows immediately from (3.1), identifying each degenerate

interval of the form $[a, a]$ with the real number a, that interval arithmetic with degenerate intervals reduces to ordinary real arithmetic. We make this identification henceforth and treat real arithmetic as a subsystem of interval arithmetic.

Associativity and commutativity with respect to addition and multiplication of intervals follows directly from the definition (3.1). In other words, if I, J, K are intervals, then the following relations hold:

$$I + (J+K) = (I+J) + K$$

$$I \cdot (J \cdot K) = (I \cdot J) \cdot K$$

$$I + J = J + I$$

$$I \cdot J = J \cdot I \ .$$

The real numbers $0, 1$ serve as identities in interval addition and interval multiplication, respectively:

$$0 + I = I + 0 = I$$

$$1 \cdot I = I \cdot 1 = I \ .$$

Inverses do not exist in general. In fact, $[a, b] - [c, d] = [a-d, b-c] = 0$ implies $a = d$ and $b = c$. Since $a \leq b$ and $c \leq d$, this means that $[a, b] - [c, d] = 0$ if and only if $a = b = c = d$. Similarly $[a, b] \cdot [c, d] = 1$ if and only if $a = b = c^{-1} = d^{-1}$.

Thus the only intervals having additive or multiplicative inverses are the real numbers themselves.

The distributive law fails, since $[1, 2] \cdot (1 - 1) = [1, 2] \cdot 0 = 0$ whereas $[1, 2] \cdot 1 + [1, 2] \cdot (-1) = [1, 2] + [-2, -1] = [-1, 1] \neq 0$.

Nevertheless, we do have the following law for any intervals I, J, K:

$$I \cdot (J+K) \subset I \cdot J + I \cdot K \tag{3.3}$$

that is, the interval $I \cdot (J+K)$ is included as a set in the interval $I \cdot J + I \cdot K$. This interesting relation, which might be called subdistributivity, follows easily from the definition (3.1).

A simple characterization of special cases in which equality holds in (3.3) has not been found, however the following cases are useful.

If t is a real number, then $t \cdot (J+K) = t \cdot J + t \cdot K$.

If $J \cdot K \geq 0$ (that is, if $x \in J \cdot K$ implies $x \geq 0$), then $I \cdot (J+K) = I \cdot J + I \cdot K$. We omit the easy proofs of these relations.

Besides (3.3) the arithmetic operations on intervals satisfy further inclusion relations which follow from the definition (3.1) :

If $I \subset K$ and $J \subset L$, then

$$(3.4) \qquad \begin{aligned} I + J &\subset K + L \\ I - J &\subset K - L \\ I \cdot J &\subset K \cdot L \\ I \div J &\subset K \div L , \quad (0 \notin L) . \end{aligned}$$

We describe this set of relations by saying that the arithmetic operations on intervals are inclusion monotonic.

The set of relations (3.4) has the important consequence that if $F(X_1, X_2, \ldots X_n)$ is a rational expression in the interval variables $X_1, X_2, \ldots X_n$, i.e. a finite combination of $X_1, \ldots X_n$ and a finite set of constant intervals with interval arithmetic operations, then

$$X'_1 \subset X_1, \; X'_2 \subset X_2, \; \ldots, \; X'_n \subset X_n$$

implies

$$(3.5) \qquad F(X'_1, X'_2, \ldots, X'_n) \subset F(X_1, X_2, \ldots, X_n) .$$

In particular, in the case that $X'_1, X'_2, \ldots, X'_n$ are real numbers and the constants in the expression for F are real numbers, then the value of $F(X'_1, X'_2, \ldots, X'_n)$ will be a real number contained in the interval $F(X_1, X_2, \ldots, X_n)$, which can be computed by a finite number of interval arithmetic operations.

Therefore we can bound the range of values of a real rational function over intervals of values for each of its arguments by evaluating a rational expression in interval arithmetic.

Furthermore, the result (3.5) implies that a digital computation carried out in "rounded" interval arithmetic (in which the machine computed end-points in the expressions in (3.2) are rounded according to the procedures discussed in section 2 above on interval numbers) produces intervals which contain the exact (or "infinite precision") results of the corresponding real arithmetic computation.

Rounded interval arithmetic is therefore, in particular, a means for the automatic determination of an upper bound to the accumulated round-off error in any digital computation.

4. Interval-valued Functions

Rational expressions which are equivalent in real arithmetic are not necessarily equivalent in interval arithmetic.

For example, consider the polynomial $p(x) = x - x^2$.

If we evaluate the expression $P_1(X) = X - X \cdot X$ using interval

arithmetic with $X = [0,1]$ we obtain

$$P_1([0,1]) = [0,1] - [0,1] \cdot [0,1] = [0,1] - [0,1] = [-1,1] .$$

Alternatively, if we evaluate the expression $P_2(X) = X \cdot (1 - X)$ using interval arithmetic with $X = [0,1]$ we obtain

$$P_2([0,1]) = [0,1] \cdot (1 - [0,1]) = [0,1] \cdot [0,1] = [0,1] .$$

Still again, if we evaluate $P_3(X) = 1/4 - (X - 1/2) \cdot (X - 1/2)$ in interval arithmetic with $X = [0,1]$ we obtain

$$P_3([0,1]) = 1/4 - [-1/2, 1/2] \cdot [-1/2, 1/2]$$

$$= 1/4 - [-1/4, 1/4] = [0, 1/2] .$$

Of course all the expressions are equivalent in real arithmetic so that for x a real number we have

$$p(x) = x - x^2 = x(1-x) = 1/4 - (x - 1/2)^2 .$$

The actual range of values of $p(x)$ for x in the interval $[0,1]$ is $[0,1/4]$ and we note that this interval, $[0,1/4]$, is, in fact, contained in each of the intervals $P_1([0,1])$, $P_2([0,1])$, $P_3([0,1])$ computed by interval arithmetic.

These three intervals are unequal on account of the failure of the distributive law.

The fact that none of them gives the exact range of values of p is due to something else.

In fact, in the case of the polynomial $y(x) = x^2$ it can be proved, [23], that there is no rational expression which in interval arithmetic computes the correct range of values, namely $[0,1]$, for $y(x)$ when x varies over the interval $[-1,1]$. The expression $X \cdot X$, for example, yields $[-1,1] \cdot [-1,1] = [-1,1]$.

The trouble is that in a numerical evaluation of an expression, the identity of variables is lost so that we must get the same value for $X \cdot Y$ with $X = [-1,1]$, $Y = [-1,1]$ as for $X \cdot X$ with $X = [-1,1]$. In other words, the direct evaluation of a rational expression in interval arithmetic in which a given variable occurs more than once may result in a wider interval than the actual range of values of the corresponding real rational function.

On the other hand computations with interval numbers need not be restricted to purely arithmetic ones.

For example, we can define an interval-valued function X^2 which takes on the values

$$X^2 = \{x^2 \mid x \in X\} .$$

And, in fact, we can program the computation of this interval-valued function using the formula:

$$[a,b]^2 = \begin{cases} [a,b] \cdot [a,b] & \text{if } 0 \notin [a,b] \\ [0, a^2] \cup [0, b^2] = [0, \max(a^2, b^2)] & \text{if } 0 \in [a,b] \end{cases}$$

More generally, if f is a real function, then using the notation $f(A)$ to represent the image under f of a set A, we have

$$f(\cup X_i) = \cup f(X_i)$$

for any collection of sets X_i in the domain of f.

In particular, if $X_i = [a_i, b_i]$ and $X = [a,b] = \cup X_i$ then

$$f([a,b]) = \cup f([a_i, b_i]) .$$

Now, if f is monotonic on each X_i, then

$$f([a_i, b_i]) = \begin{cases} [f(a_i), f(b_i)] & \text{if } f \text{ increases on } X_i \\ [f(b_i), f(a_i)] & \text{if } f \text{ decreases on } X_i \end{cases}$$

These considerations are of use in programming extensions of real functions, for example the sine and cosine functions, to interval-valued functions on intervals.

Using a programmed approximation of known accuracy for $\sin x$, when x is a real machine number, and using the fact that the sine function is piecewise monotonic with relative maxima and minima at known locations, we can write a program which computes an interval $SIN(X)$ for an arbitrary interval $X = [a,b]$ with a, b machine numbers such that

$$SIN(X) \supset \{\sin x \mid x \in X\} .$$

With care, this can be done in such a way that the width of the interval $SIN(X)$ is only slightly greater than the width of the interval $\{\sin x \mid x \in X\}$.

Interval-valued extensions of all the commonly used elementary function "subroutines" can be obtained in various ways.

Any given real rational function is piecewise monotonic; however, the relative maxima and minima occur at roots of algebraic equations whose locations may not be known in advance.

On account of the inclusion relation (3.5) we know that a straight forward evaluation in interval arithmetic of a rational expression will produce, if division by an interval containing zero does not occur, an interval containing the range of values of the corresponding real rational function for real arguments ranging over the argument intervals used.

On the other hand, we saw in the examples above that the width of the containing interval thus obtained may be greater than the width of the exact interval of the range of real values of the real rational function.

We will show now that the exact range of values can be approached arbitrarily closely by a finite union of intervals.

Suppose F is a rational interval function with $F(X)$ defined for all intervals X contained in some interval A by a particular interval arithmetic expression with real coefficients in the interval variable X. The values of F on real numbers will be real, (with the identification of degenerate intervals and real numbers agreed upon, above in section 3). Denote the real rational function by f so that

$$f(x) = F([x,x]), \quad x \in A .$$

Denote the exact range of values of f on X by $f(X)$; thus $f(X) = \{f(x) \mid x \in X\}$. Denote the width of an interval X by $w(X)$, thus $w([a,b]) = b - a$.

Theorem 1. There is a positive real number K depending on F and A but independent of the method of subdivision of the interval X such that if X is the union of subintervals X_i, then

$$f(X) \subset \bigcup_{i=1}^{n} F(X_i)$$

and

$$w(\bigcup_{i=1}^{n} F(X_i)) \leq w(f(X)) + K \max_i w(X_i) .$$

Proof: The first part of the theorem follows from the inclusion relation (3.5) and the fact that

$$f(X) = \bigcup_{i=1}^{n} f(X_i) .$$

To prove the second part we need to show that if

$$y_1 \in \bigcup_{i=1}^{n} F(X_i)$$

then there is a

$$y_2 \in f(X)$$

such that

$$|y_1 - y_2| \leq K \max_i w(X_i) \ .$$

In fact if

$$y_1 \in F(X_i)$$

then we will show that for every $y_2 \in f(X_i)$ we have

$$|y_1 - y_2| \leq K\, w(X_i) \ .$$

In the expression for $F(X)$, the variable X occurs only a finite number of times, say J times (possibly zero). In each occurrence substitute a new variable $X^{(j)}$, $j = 1, 2, \ldots, J$.

(For example, in the expression $X \cdot X$ substitute $X^{(1)} \cdot X^{(2)}$).

Call the new expression $H(X^{(1)}, X^{(2)}, \ldots, X^{(J)})$ thus $F(X) = H(X, X, \ldots, X)$. And for real x, $H(x, x, \ldots, x) = F(x) = f(x)$. But the expression $H(X^{(1)}, X^{(2)}, \ldots, X^{(J)})$ defines a real rational function for real $x^{(1)}, \ldots, x^{(J)}$ and there is a Lipschitz constant K for $x, x^{(1)}, \ldots, x^{(J)}$ in A such that

$$|H(x^{(1)}, \ldots, x^{(J)}) - H(x, x, \ldots, x)| \leq K \max_j |x^{(j)} - x| \ .$$

Now if $y_1 \in F(X_i)$, then $y_1 = H(x^{(1)}, \ldots, x^{(J)})$ for some set of values $x^{(1)}, x^{(2)}, \ldots, x^{(J)}$ in X_i. But for every such set and every x in X_i we have $|y_1 - H(x, x, \ldots, x)| \leq K\, w(X_i)$. Recall that $H(x, x, \ldots, x) = f(x)$; and $y_2 \in f(X)$ is the same as $y_2 = f(x)$ for some $x \in X$. Therefore, for every $y_2 \in f(X_i)$, it follows that $|y_1 - y_2| \leq K\, w(X_i)$.

This completes the proof of Theorem 1.

More general theorems of this type for rational interval functions in any finite number of interval variables with interval

coefficients also have been proved, [21], [23].

Along with the notation $w([a,b]) = b - a$ for the width of an interval, it is also useful to define the "magnitude" of an interval by $|[a,b]| = \max(|a|, |b|)$. The following relations between the interval arithmetic operations and the widths and "magnitudes" of intervals are easily demonstrated.

For positive real numbers a, b and any intervals I, J

$$w(aI + bJ) = a\,w(I) + b\,w(J)$$

$$w(IJ) \leq |I|\,w(J) + |J|\,w(I)$$

$$|I + J| \leq |I| + |J|$$

$$w(-J) = w(J), \quad \text{where } -[a,b] = [-b, -a] .$$

For any real number a, and any interval I

$$|aI| = |a|\,|I|$$

$$w(aI) = |a|\,w(I) .$$

And for any interval I which does not contain the real number zero,

$$w(1 \div I) \leq |1 \div I|^2\, w(I) .$$

As a note of caution, we point out that

$$|1 \div [a,b]| = |[1/b, 1/a]| = \begin{cases} 1/a & \text{if } 0 < a \leq b \\ -1/b & \text{if } a \leq b < 0 \end{cases}$$

whereas

$$1 \div |[a,b]| = \begin{cases} 1/b & \text{if } 0 < a \leq b \\ -1/a & \text{if } a \leq b < 0 \end{cases}$$

Therefore $a \neq b$ implies

$$|1 \div [a,b]| \neq 1 \div |[a,b]| .$$

5. Interval Contractions

In this section we will illustrate the application of the above considerations to the study of some iterative computations with interval numbers.

We will give examples of the computation of sequences of intervals of decreasing width containing and converging to an irrational real number.

Consider first the rational interval function F defined for $X \subset [1, 2]$ by $F(X) = 1 + \frac{1}{1+X} \equiv 1 + \{1 \div (1 + X)\}$. We have

$$\begin{aligned} F([1, 2]) &= 1 + \{1 \div (1 + [1, 2])\} \\ &= 1 + \{1 \div [2, 3]\} \\ &= 1 + [1/3, 1/2] \\ &= [4/3, 3/2] \end{aligned}$$

so that $F([1, 2] \subset [1, 2]$.

Define a sequence of intervals for $n = 0, 1, 2, \ldots$, by

$$X_{n+1} = F(X_n), \quad X_0 = [1, 2] .$$

By the inclusion relation (3.5) we know that for $X' \subset X$ we have $F(X') \subset F(X)$ therefore, since we have already seen that $X_1 = F(X_0) \subset X_0$, it follows by induction on n that $X_{n+1} \subset X_n$. Furthermore, from the relations given at the end of the last section, we have

$$w(F(X)) \leq |1 \div (1+X)|^2 \, w(X)$$

and for $X \subset [1, 2]$ this implies $w(F(X)) \leq (1/4)\, w(X)$.

Applying this inequality to the sequence $\{X_n\}$ we find that $w(X_n) \leq (1/4)^n$. Therefore $\{X_n\}$ is a nested sequence of intervals of widths converging to zero and the limit of the sequence is the number $x = \sqrt{2}$ satisfying

$$F(x) = x = 1 + \frac{1}{1+x}$$

or $x^2 = 2$.

Using three significant decimal digit interval arithmetic we obtain the sequence

$$\begin{aligned} X_0 &= [1, 2] \\ X_1 &= [1.33, 1.50] \\ X_2 &= [1.40, 1.43] \\ X_3 &= X_4 = \ldots = X_n = [1.41, 1.42], \quad n \geq 3 . \end{aligned}$$

Another way of looking at the sequence is that it generates the interval-valued partial continued fractions

$$1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{\ddots + \cfrac{1}{[2,3]}}}}$$

each of which contains the number $\sqrt{2}$. By $1/X$ we mean, of course, $1 \div X$.

Next, consider the interval mapping G defined for $Y \subset [1,2]$ by $G(Y) = (1/2)Y + 1/Y$. Again $\sqrt{2}$ is a fixed point of G.

We have

$$\begin{aligned} G([1,2]) &= 1/2\,[1,2] + 1/[1,2] \\ &= [1/2, 1] + [1/2, 1] \\ &= [1,2] \end{aligned}$$

Therefore the sequence $\{Y_k\}$ defined by $Y_{k+1} = G(Y_k)$ repeats with Y_0 chosen as $[1,2]$. And with $Y_0 = [1.4, 1.5]$, we obtain $Y_1 = G([1.4, 1.5]) = [1.36.., 1.46..]$ which is not contained in Y_0.

On the other hand, using a decomposition of the intervals we can obtain a nested sequence converging to $\sqrt{2}$ using unions of a fixed number of subintervals. For $Y = [y_1, y_2]$, define, for positive integers n,

$$\Delta_i^n Y = y_1 + [i-1, i] \left\{\frac{y_2 - y_1}{n}\right\}, \quad i = 1, 2, \ldots, n,$$

then

$$Y = \bigcup_{i=1}^{n} \Delta_i^n Y$$

and define $G^{(n)}(Y) = \bigcup_{i=1}^{n} G(\Delta_i^n Y)$

where $G(Y)$ is defined as above.

By the inclusion relation (3.5) we have $G^{(n)}(Y) \subset G(Y)$.

It can be shown, [23], that for each $n \geq 2$ the sequence $\{Y_k^{(n)}\}$ $k = 1, 2, \ldots$ defined by $Y_{k+1}^{(n)} = G^{(n)}(Y_k^{(n)})$ converges to $\sqrt{2}$, with $Y_0^{(n)} = [1,2] \supset Y_1^{(n)} \supset Y_2^{(n)} \ldots \supset \sqrt{2}$.

We conclude this section with a final example of a sequence

of contracting intervals -- again converging to $\sqrt{2}$. Consider the mapping $f(y) = y^2 - 2$. According to the mean value theorem $f(y) = f(x) + f'(x + \theta(y - x))(y - x)$ for some $\theta \in [0,1]$.

If x is the positive zero of $f(x)$, namely $\sqrt{2}$, then

$$x = y - \frac{f(y)}{f'(x + \theta(y - x))} = y + \frac{1 - 1/2\, y^2}{x + \theta(y - x)}$$

for some $\theta \in [0,1]$, therefore (by 3.5)

$$x \in y + (1 - 1/2\, y^2) \div (x + [0,1] \quad (y - x))$$

provided $0 \notin x + [0,1]\ (y - x)$ (since $x = \sqrt{2}$, this amounts to the restriction $y > 0$).

Denote by mY the midpoint of the interval Y; thus $m[a,b] = (a + b)/2$. Define the sequence of intervals $\{Y_n\}$ by

$$Y_0 = [1, 2]$$

$$Y_{n+1} = mY_n + (1 - 1/2(mY_n)^2) \div Y_n .$$

Thus

$$Y_1 = 3/2 + (1 - 1/2(3/2)^2 \div [1, 2] = [1.375,\ 1.4375]$$

$$Y_2 = [1.41406 \ldots,\ 1.41441 \ldots]$$

$$Y_3 = [1.414213559 \ldots,\ 1.414213563 \ldots]$$

and the sequence $\{Y_n\}$ is a nested sequence of intervals, [23], contracting to the real number $\sqrt{2}$.

6. A Metric Topology for Interval Numbers

The introduction of a metric or distance function for interval numbers provides for the concept of a continuous interval function and is a valuable aid both to the analysis of interval functions and to the geometric intuition concerning the closeness of two intervals.

Denote by $\mathcal{I}$ the set of closed real intervals

$$[a,b], \quad a \leq b .$$

We make $\mathcal{I}$ into a metric space with the distance function $P([a,b], [c,d]) = \max(|a - c|, b - d|)$.

Notice that for degenerate intervals $[a,a]$, $[b,b]$ we have

$P([a,a],[b,b]) = |a-b|$.

Thus our metric P is consistent with our identification of the degenerate interval $[a,a]$ with the real number a; the real line may be regarded as a subspace of the metric space $(\mathcal{I}, P)$.

It has been shown, [21], [23], [30], [31], that the arithmetic operations in $\mathcal{I}$ are continuous except, of course, for division by intervals containing zero. It follows from this that the rational interval functions are continuous.

The set of interval numbers $[x,y]$ may be pictured as the half-plane of points (x,y), $x \leq y$, above the diagonal $y = x$, with the diagonal itself corresponding to the real numbers (fig. 1).

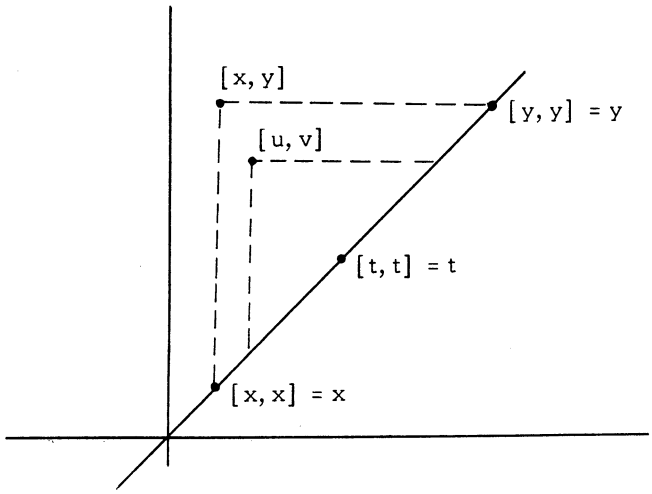

figure 1

In figure 1, the interval number $[u,v]$ is seen to be contained in the interval number $[x,y]$, since $x < u < v < y$. Furthermore we see that the interval $[u,v]$ contains all real numbers t, $u \leq t \leq v$, i.e., points on the diagonal segment intercepted by the horizontal and vertical lines through $[u,v]$.

Two interval numbers are close with respect to our metric P if the corresponding points are close in the diagram.

7. Interval Integrals

Suppose F is a rational interval function with real coefficients defined for $X \subset A$. Then $f(x) = F([x,x])$ is a bounded real rational function on the interval A.

If $X = [a,b] \subset A$, then $f(x) \in F(X)$ for all $x \in X$, and by the mean value theorem, we have

$$\int_a^b f(x)\,dx = f(a + \theta(b-a))(b-a) \quad ,$$

for some $\theta \in [0,1]$. Therefore,

$$\int_a^b f(x)\,dx \in F(X)(b-a) \quad . \tag{7.1}$$

Now suppose $Y = [a,y] \subset A$; define

$$Y_i^{(n)} = a + [i-1,i]\,\frac{(y-a)}{n} \quad ,$$

then $Y_i^{(n)} \subset A$, and

$$Y = \bigcup_{i=1}^{n} Y_i^{(n)} \quad .$$

By the additivity of the integral, we have

$$\int_a^y f(x)\,dx \in \sum_{i=1}^{n} F(Y_i^{(n)})\,\frac{(y-a)}{n} \quad \text{[interval sum]} \quad . \tag{7.2}$$

More generally, if $Y_i^{(n)}$, $i = 1, 2, \ldots, n$, is a collection of intervals such that

$$Y_i^{(n)} \subset A,$$

$$Y = \bigcup_{i=1}^{n} Y_i^{(n)} = [a,y] \quad ,$$

and

$$w(Y) = \sum_{i=1}^{n} w(Y_i^{(n)}) = y - a$$

then

$$\int_a^y f(x)\,dx \in \sum_{i=1}^{n} F(Y_i^{(n)})\,w(Y_i^{(n)}) \quad . \tag{7.3}$$

The intervals $Y_i^{(n)}$ do not all have to have the same width. From Theorem 1 it follows that there exists a K such that for $X \subset A$, we have $w(F(X)) \leq K\, w(X)$, and therefore

$$(7.4) \qquad w\left(\sum_{i=1}^{n} F(Y^{(n)})\, w(Y_i^{(n)})\right) \leq (y - a) K \max_i w(Y_i^{(n)}), \quad i = 1, \ldots, n,$$

and we have proved the following:

Theorem 2.

Define

$$I_n(y) = \sum_{i=1}^{n} F(Y_i^{(n)})\, w(Y_i^{(n)}) ,$$

then

$$I_n(y) = \int_a^y f(x)\, dx + E$$

with $0 \in E$, and

$$w(E) \leq (y - a)\, K \max w(Y_i^{(n)}) .$$

For an illustrative example, denote the natural logarithm of y by $\log y$, and consider

$$\log y = \int_1^y \frac{dx}{x} , \quad y > 1 ,$$

and take $F(Y) = 1/Y$ for $Y \geq 1$ (i.e., $y \in Y$ implies $y \geq 1$).

Let

$$Y_i^{(n)} = 1 + [i - 1, i]\frac{(y-1)}{n} , \quad i = 1, 2, \ldots, n ,$$

then

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

If $a \geq 1$, then

$$w(F([a, b])) = w[1/b, 1/a] = \frac{b - a}{ab} \leq w([a, b]) ,$$

so we can take $K = 1$ in Theorem 2 and obtain for $n = 1, 2, \ldots$,

$$\log y \in I_n(y) \quad \text{and} \quad w(I_n(Y)) \leq \frac{y - 1}{n} .$$

The interval-valued functions

$$Y_i^{(n)}(y), \; I_n(y) ,$$

of the real variable y occurring in Theorem 2 can be extended to rational interval functions on $Y' \geq 1$ in the following way:

$$Y_i^{(n)}(Y') = 1 + [i - 1, i] \frac{Y' - 1}{n} ,$$

$$I_n(Y') = \sum_{i=1}^{n} F(Y_i^{(n)}(Y')) w(Y_i^{(n)}(Y')) .$$

Clearly, $\{\log y \mid y \in Y'\} \subset I_n(Y')$ and

$$\int_{Y'} \log y \, dy \in I_n(Y') \, w(Y') .$$

Suppose $Y' = [y, y']$ with $y \geq 1$. Let

$$Y_j^{(m)} = y + [j - 1, j] \frac{(y' - y)}{m} ,$$

then

$$Y' = \bigcup_{j=1}^{m} Y_j^{(m)}$$

and

$$\sum_{j=1}^{m} w(Y_j^{(m)}) = w(Y') .$$

Now $I_n(Y')$ is a rational interval function of Y', so by Theorem 1 there exists a positive real number K such that

$$\bigcup_{j=1}^{m} I_n(Y^{(m)}) = \overline{I}_n(Y') + E_m$$

with $0 \in E_m$ and $w(E_m) \leq K \frac{(y' - y)}{m}$. Since $w(I_n(y)) \leq \frac{y - 1}{n}$, and $\log y \in I_n(y)$, we have

$$\overline{I}_n(Y') = \bigcup_{y \in Y'} I_n(y) \subset \{\log y \mid y \in Y'\} + [-1,1]\frac{y'-1}{n}.$$

Therefore

$$\bigcup_{j=1}^{m} I_n(Y_j^{(m)}) = \{\log y \mid y \in Y'\} + E_{n,m},$$

with $0 \in E_{n,m}$, $w(E_{n,m}) \leq \frac{K\, w(Y')}{m} + \frac{2(y'-1)}{n}$. It is also easily shown that there are positive real numbers K, K', such that

$$\sum_{j=1}^{m} I_n(Y_j^{(m)})\, w(Y_j^{(m)}) = \int_{Y'} \log y\, dy + E'_{n,m},$$

with $0 \in E'_{n,m}$ and

$$w(E'_{n,m}) \leqq \frac{K\, w(Y') + K'\{w(Y')\}^2}{m} + \frac{2\, w(Y')(y'-1)}{n}.$$

Thus by iterating and composing the processes of taking unions and interval integration we can obtain, by finite computations with intervals, sequences of intervals containing and converging to the range of values and the integrals of real valued functions such as the logarithm which are not themselves rational but which are integrals or even iterated integrals of rational functions.

We consider now some more rapidly convergent procedures for bounding real integrals with interval computations.

Suppose $F^{(0)}, F^{(1)}, \ldots, F^{(k)}$ are rational interval functions defined for $X \subset A$ such that the corresponding real valued rational functions $\{f^{(r)}\}$ are the successive derivatives up to order k of an ordinary rational function with real coefficients, defined by $F^{(0)}(x) = f^{(0)}(x) = f(x)$. We know that f is bounded on A because $x \in A$ implies $f(x) \in F^{(0)}(A)$. Consider the real integral

$$\int_a^b f(x)\, dx$$

with $[a,b] \subset A$.

Subdivide the interval $[a,b]$ so that

$$[a,b] = \bigcup_{i=1}^{n} X_i$$

with

$$\sum_{i=1}^{n} w(X_i) = b - a$$

and write $X_i = [x_{i-1}, x_i]$ and $f = f^{(0)}$.

The Taylor theorem with remainder asserts that for each $t \in [0, w(X_i)]$

$$f(x_{i-1} + t) = f^{(0)}(x_{i-1}) + f^{(1)}(x_{i-1})t + \ldots + \frac{f^{(k-1)}(x_{i-1})}{(k-1)!}t^{k-1} + R_{i-1}^{(k)}(t)$$

with

$$R_{i-1}^{(k)}(t) = \frac{1}{k!} f^{(k)}(x_{i-1} + \theta_t t)t^k$$

for some $\theta_t \in [0,1]$.

Now

$$\int_{X_i} f(x)\,dx = \sum_{r=0}^{k-1} \frac{f^{(r)}(x_{i-1})}{r!} \int_0^{w(X_i)} t^r\,dt + \int_0^{w(X_i)} R_{i-1}^{(k)}(t)\,dt \ .$$

The last integral exists since all the others do.

We can write

$$\int_0^h g(t)\,t^k\,dt = \frac{1}{k+1} \int_0^h g(t)\,d(t^{k+1}) \ ;$$

therefore, using (7.1), we obtain

$$\int_0^{w(X_i)} R_{i-1}^{(k)}(t)\,dt = \frac{1}{k!} \int_0^{w(X_i)} f^{(k)}(X_{i-1}+\theta_t t)t^k\,dt \in \frac{1}{(k+1)!} F^{(k)}(X_i)\{w(X_i)\}^{k+1}$$

Since

$$\int_0^{w(X_i)} t^r\,dt = \frac{1}{r+1}\{w(X_i)\}^{r+1} \ ,$$

we finally have the result that

$$\int_a^b f(x)\,dx \in \sum_{i=1}^{n} \sum_{r=0}^{k-1}{}' \frac{f^{(r)}(x_{i-1})}{(r+1)!} \{w(X_i)\}^{r+1} + E_{n,k}$$

with

$$E_{n,k} = \frac{1}{(k+1)!} \sum_{i=1}^{n} F^{(k)}(X_i) \{w(X_i)\}^{k+1} .$$

Since there is a K_k such that for all $X_i \subset A$,

$$w\left(F^{(k)}(X_i)\right) \leqq K_k\, w(X_i) ,$$

it follows that

$$w(E_{n,k}) \leqq \frac{K_k}{(k+1)!} (b-a) \max_{i=1,\ldots,n} \{w(X_i)\}^{k+1} .$$

Now define $I_{n,k}, (n, k \geqq 1)$ by

$$(7.5) \quad I_{n,k} = \sum_{i=1}^{n} \sum_{r=0}^{k-1} \frac{F^{(r)}(x_{i-1})}{(r+1)!} \{w(X_i)\}^{r+1} + \frac{1}{(k+1)!} \sum_{i=1}^{n} F^{(k)}(X_i) \{w(X_i)\}^{k+1} .$$

We have proved that

Theorem 3. $$\int_a^b f(x)\,dx \in I_{n,k}, \quad n, k \geqq 1$$

and if $w(X_i) \leqq h$ for $i = 1, \ldots, n$, then

$$(7.6) \quad w(I_{n,k}) \leqq \frac{K_k}{(k+1)!} (b-a) h^{k+1} .$$

The formula (7.5) gives a $(k+1)^{st}$ order method in the sense of (7.6) for each positive integer k.

In case $k = 0$, delete the double sum on the right hand side of (7.5) and the first order method expressed by Theorem 2 results.

For an example, consider

$$\int_1^2 \frac{dx}{x}.$$

Let $X_i = 1 + [i-1, i]\frac{1}{n}$, $(i = 1, \ldots, n)$. We have $f^{(0)}(x) = f(x) = \frac{1}{x}$, so

$$f^{(r)}(x) = \frac{(-1)^r r!}{x^{r+1}}, \qquad r = 0, 1, 2, \ldots$$

Now take

$$F^{(r)}(X) = \frac{(-1)^r r!}{X^{r+1}}, \qquad r = 0, 1, 2, \ldots$$

then

$$w\left(F^{(r)}(X)\right) = r!\, w\left(\frac{1}{X^{r+1}}\right).$$

If $X = [a, b] \subset [1, 2]$, then

$$\frac{1}{[a, b]^{r+1}} = \left[\frac{1}{b^{r+1}}, \frac{1}{a^{r+1}}\right]$$

and

$$w\left(\frac{1}{X^{r+1}}\right) = \frac{1}{a^{r+1}} - \frac{1}{b^{r+1}} = (b-a)\frac{(b^r + \ldots + a^r)}{a^{r+1} b^{r+1}}$$

$$\leqq \frac{(r+1)!}{b\, a^{r+1}}\, w(X)$$

$$\leqq (r+1)\, w(X)$$

Thus we can use $K_k = (k+1)!$, $b - a = 1$, $h = \frac{1}{n}$ in (7.6) to obtain

$$\int_1^2 \frac{dx}{x} \in I_{n,k}$$

with

$$(7.7) \qquad w(I_{n,k}) \leqq \left(\frac{1}{n}\right)^k .$$

where

$$(7.8) \quad I_{n,k} = \sum_{i=1}^{n} \sum_{r=0}^{n-1} \frac{(-1)^r}{r+1} \left(1 + \frac{i-1}{n}\right)^{-r-1} \left(\frac{1}{n}\right)^{r+1} + \frac{1}{k+1} \sum_{i=1}^{n} (-1)^k \left(1 + [i-1, i]\frac{1}{n}\right)^{-k-1} \left(\frac{1}{n}\right)^{k+1} .$$

Call $y_i = \frac{1}{n+i-1}$, then (7.8) can be rewritten as

$$(7.9) \quad I_{n,k} = \sum_{i=1}^{n} \left\{ y_i \left(1 + y_i \left(-\frac{1}{2} + \ldots + y_i \left(\frac{(-1)^{k-1}}{k}\right) \ldots\right)\right)\right\} + \frac{(-1)^k}{k+1} \sum_{i=1}^{n} [n+i-1,\ n+i]^{-k-1} .$$

We now give a heuristic discussion of the "efficiency" of this formula.

Using (7.9), the computation of $I_{n,k}$ requires roughly $3kn$ additions and multiplications of real numbers (ignoring about $n+2$ divisions). Looking at the bound (7.7), suppose we wish to make

$$\left(\frac{1}{n}\right)^k = \epsilon$$

then

$$k = \frac{\log 1/\epsilon}{\log n} .$$

The quantity $C_{n,k} = 3kn$ measures the amount of computation required to evaluate $I_{n,k}$. In order to achieve $w(I_{n,k}) \leqq \epsilon$ it is sufficient to use any positive integer n together with k_n, the smallest integer satisfying

$$k_n \geq \frac{\log 1/\epsilon}{\log n} .$$

Then the amount of computation required will be $C_{n,k_n} = 3k_n n$ or very nearly

$$C(n) = \frac{3n}{\log n} \log \frac{1}{\epsilon} .$$

The function $C(n)$ has a minimum at $n = 3$ for positive integers n, so the most efficient choice of k, n indicated by this argument is $n = 3$, and $k_n = k$ the smallest integer satisfying

$$k_n \geqq \frac{\ln \frac{1}{\epsilon}}{\ln 3} \ ;$$

in this case we find that the amount of computation required for

$$\left(\frac{1}{n}\right)^k = \epsilon$$

if very nearly

$$C(3) = (8.19 \ldots) \log \frac{1}{\epsilon} \ .$$

If $\epsilon = 10^{-10}$ for example, we choose $n = 3$, $k = 21 \sim \frac{\log 10^{10}}{\log 3} = 20.9\ldots$ and using (7.9) to compute $I_{3,\,21}$ we would have $C_{3,\,21} = 189$,

$$\begin{aligned} w(I_{3,\,21}) &= w\left(\frac{1}{22} \sum_{i=1}^{3} [3 + i - 1,\ 3 + i]^{22}\right) \\ &= \frac{1}{22}\{w([3,4]^{-22}) + w([4,5]^{-22}) + w([5,6]^{-22})\} \\ &= \frac{1}{22}\{\left(\frac{1}{3}\right)^{22} - \left(\frac{1}{6}\right)^{22}\} \\ &< 10^{-10} \end{aligned}$$

as claimed.

Suppose we arbitrarily choose a value of k, say $k = 4$. Then to guarantee $w(I_{n,\,4}) \leqq 10^{-10}$ with this "4^{th} order" method, we need to take n according to (7.7) such that

$$\left(\frac{1}{n}\right)^4 \leqq 10^{-10} \ ,$$

or $n \geqq 317$ and in this case $C_{317,\,4} = 3 \cdot 317 \cdot 4 = 3804$. Or, in other words, this choice requires about 20 times as much computation as our "most efficient" choice.

The method defined by (7.5) was based on a local expansion of the integrand $f(x)$ in Taylor's series.

It has become traditional to discard numerical methods based on direct expansions in Taylor series as impractical because of the difficulty in computing the Taylor coefficients. We no longer subscribe to this point of view; in fact, in section 9 below we present a practical technique for the use of Taylor expansions on a digital computer.

We conclude this section with an interval version of Gaussian quadrature. We write

$$[a, b] = \bigcup_{i=1}^{n} X_i$$

with

$$\sum_{i=1}^{n} w(X_i) = b - a$$

with the same assumptions on f as in (7.5). Then $X_i = [x_{i-1}, x_i]$ and the Gaussian method has the form

$$(7.10) \qquad \int_a^b f(x)\,dx = \sum_{i=1}^{n} \left\{ w(X_i) \sum_{r=1}^{k} g_r f(x_{i-1} + u_r(w(X_i))) \right\} + E_{n,k}$$

where

$$(7.11) \qquad E_{n,k} = \frac{(k!)^4}{[(2k)!]^3 (2k+1)} \sum_{i=1}^{n} \{w(X_i)\}^{2k+1} f^{(2k)}(\xi_i)$$

for some $\xi_i \in X_i$, $i = 1, 2, \ldots, n$.

The numbers g_r and u_r $(r = 1, 2, \ldots, k)$ are the weights and argument spacings of the Gauss "k-point formula" [19]. They are associated with the zeros of the Legendre polynomials

$$P_k(t) = \frac{d^k}{dt^k} (t^2 - 1)^k$$

and are tabulated to 15 decimal place accuracy for $k = 1, \ldots, 16$ in [18].

Using Stirlings inequalities for $n!$ we find that for positive integer values of k:

$$\frac{2k}{2k+1} 2\pi \left(\frac{1}{4}\right)^{2k+1} < \frac{[k!]^4}{[(2k)!]^2 (2k+1)} < 2\pi \left(\frac{1}{4}\right)^{2k+1}$$

Since $f^{(2k)}(\xi_i) \in F^{(2k)}(X_i)$ for a rational interval function $F^{(2k)}$ with real restriction $f^{(2k)}$, we can write (using (7.11))

$$E_{n,k} \in 2\pi[\frac{2k}{2k+1}, 1] \sum_{i=1}^{n} \left(\frac{w(X_i)}{4}\right)^{2k+1} \frac{F^{(2k)}(X_i)}{(2k)!} .$$

We now define an interval version of the Gaussian method by the formula

$$(7.12)\quad I^{G}_{n,k} = \sum_{i=1}^{n} w(X_i) \sum_{r=1}^{k} g_r f(x_{i-1} + u_r w(X_i)) + 2\pi[\frac{2k}{2k+1}, 1] \sum_{i=1}^{n} \left(\frac{w(X_i)}{4}\right)^{2k+1} \frac{F^{(2k)}(X_i)}{(2k)!}$$

For this method we have

$$\int_a^b f(x)\,dx \in I^{G}_{n,k} ;$$

for $X \subset A$, there is a K_{2k} such that $w(F^{(2k)}(X)) \leqq K_{2k}\, w(X)$ and for $h = \max_{i=1,2,\ldots,n} w(X_i)$ we have

Theorem 4

$$w(I^{G}_{n,k}) \leqq \frac{2\pi}{(2k)!}(b-a)\{\frac{h}{4}K_{2k} + \frac{1}{4(2k+1)}|F^{(2k)}([a,b])|\}(\frac{h}{4})^{2k} .$$

Proof: Recall that $w(AB) \leqq |A|\, w(B) + |B|\, w(A)$ and $w(aA + bB) = |a|\, w(A) + |b|\, w(B)$.

Thus

$$w(I^{G}_{n,k}) \leqq 2\pi \left\{ (\frac{h}{4})^{2k+1} \frac{K_{2k}(b-a)}{(2k)!} + \frac{1}{(2k+1)(2k)!} \sum_{i=1}^{n} \left(\frac{w(X_i)}{4}\right)^{2k+1} |F^{(2k)}(X_i)| \right\}.$$

Now

$$\sum_{i=1}^{n}\left(\frac{w(X_i)}{4}\right)^{2k+1}|F^{(2k)}(X_i)|\leq\frac{1}{4}\left(\frac{h}{4}\right)^{2k}\sum_{i=1}^{n}|F^{(2k)}(X_i)|w(X_i)$$

and

$$\sum_{i=1}^{n}|F^{(2k)}(X_i)|w(X_i)\leq|F^{(2k)}([a,b])|(b-a)\ .$$

Putting these inequalities together we obtain Theorem 4.

Returning to our example

$$\int_1^2\frac{dx}{x}\ ,$$

put

$$F^{(2k)}(X)=\frac{(-1)^{2k}(2k)!}{X^{2k+1}}$$

as before. We can take $K_{2k}=(2k+1)!$ and by direct computation we find that

$$|F^{(2k)}([1,2])|=(2k)!\ .$$

So with $w(X_i)=\frac{1}{n}$ we have

$$\int_1^2\frac{dx}{x}\in I^G_{n,k}$$

with (7.12) becoming

$$(7.13)\quad I^G_{n,k}=\sum_{i=1}^{n}\frac{1}{n}\sum_{r=1}^{k}g_r\,\frac{1}{1+\frac{i-1}{n}+u_r\frac{1}{n}}+2\pi\left[\frac{2k}{2k+1},1\right]\sum_{i=1}^{n}\left(\frac{1}{4n}\right)^{2k+1}\frac{(-1)^{2k}}{\left[1+\frac{i-1}{n},1+\frac{i}{n}\right]^{2k+1}}$$

and (using Theorem 4), we obtain

$$(7.14)\qquad w(I^G_{n,k})\leq 2\pi\left\{\frac{1}{4n}(2k+1)+\frac{1}{4(2k+1)}\right\}\left(\frac{1}{4n}\right)^{2k}\ .$$

Counting a division as 3 multiplications, the number of multiplications and additions to evaluate $I^G_{n,k}$ by (7.13) is roughly (ignoring n additions)

$$C^G_{n,k} = (4k + 8)n .$$

In order to achieve $w(I_{n,k}) \leqq \epsilon$ it is sufficient to take n, k positive integers such that

$$2\pi \{\frac{1}{4n}(2k + 1) + \frac{1}{4(2k + 1)}\} (\frac{1}{4n})^{2k} \leqq \epsilon .$$

If $\epsilon = 10^{-10}$, we can choose $n = 1$ and $k = 10$, in which case $C^G_{1,10} = 48$. This is evidently the most efficient choice of n, k in this example; if $w(I^G_{n,k}) \leqq 10^{-10}$ and $C^G_{n,k} \leqq 48$, then $n = 1$, $k = 10$. Recall that our best choice of n, k for $I_{n,k}$ with the same value for ϵ was $n = 3$, $k = 21$ in which case $C_{3,21}$ was 189.

In other words, it takes about a fourth as many arithmetic operations to evaluate

$$\int_1^2 \frac{dx}{x}$$

using $I^G_{n,k}$ as it does using $I_{n,k}$ to achieve guaranteed ten decimal place accuracy.

8. The Initial Value Problem in Ordinary Differential Equations

In this section we are concerned with computing intervals containing values of the solution to the system of first order ordinary differential equations

$$\frac{dy_j}{dx} = f_j(x, y_1, \ldots, y_m), \quad j = 1, \ldots, m , \tag{8.1}$$

satisfying the initial conditions

$$y_j(x_o) = y_o, \quad j = 1, \ldots, m . \tag{8.2}$$

For brevity, we will sometimes use the vector notation y for $(y_1, \ldots, y_m)$ and f for $(f_1, \ldots, f_m)$. For example, we can write (8.1) in the simpler form

$$\frac{dy}{dx} = f(x, y) \tag{8.3}$$

and (8.2) can be written $y(x_o) = y_o$. We will use the metric $|y - z| = \max \{|y_1 - z_1|, \ldots, |y_m - z_m|\}$ for m-dimensional vectors y, z, f, etc.

It is well known [16] that when f is continuous on $D_f \supset B = [x_o, a] \otimes B_1 \otimes B_2 \ldots \otimes B_m$ with $a > x_o$ and y_{jo} in the non-empty interior of the interval B_j, $(j = 1, \ldots, m)$, and when f satisfies a Lipschitz condition on D_f

$$|f(x, y_1) - f(x, y_2)| \leqq K_f |y_1 - y_2| \tag{8.4}$$

for some non-negative real number K_f, then there exists exactly one solution to (8.1) and (8.2) in $D_f^* \subset D_f$ with $D_f^* = [x_o, x^*] \otimes B_1, \ldots, \otimes B_m$ for x^* such that $(x, y) \in B$ implies

$$y_{jo} + (x^* - x_o) f_j(x, y) \in B_j \qquad (j = 1, 2, \ldots, m) .$$

Denote the set of closed real intervals by $\mathcal{I}$ and the set of closed subintervals of $A \in \mathcal{I}$ by $\mathcal{I}_A$. We will suppose throughout this section that $F_1, \ldots, F_m$ are interval valued functions on the domain $D_F = \mathcal{I}_{[x_o, a]} \otimes \mathcal{I}_{B_1}, \ldots, \otimes \mathcal{I}_{B_m}$ satisfying the following conditions for $j = 1, 2, \ldots, m$:

1) F_j is continuous and F_j restricted to $B = [x_o, a] \otimes B_1, \ldots, \otimes B_m$ is a real valued function f_j, i.e., $F_j(x, y_1, \ldots, y_m) = f_j(x, y_1, \ldots, y_m)$ for $(x, y_1, \ldots, y_m) \in B$;

2) F_j is inclusion monotonic, i.e., $X' \subset X$, $Y'_1 \subset Y_1, \ldots, Y'_m \subset Y_m$ implies $F_j(X', Y'_1, \ldots, Y'_m) \subset F_j(X, Y_1, \ldots, Y_m)$

3) There is a real number K_F such that $w(F_j(X, Y_1, \ldots, Y_m)) \leqq K_F \max \{w(x), w(Y_1), \ldots, w(Y_m)\}$

Notice that in case F is a rational interval function on D_F with real restriction f on B then the conditions 1), 2), 3) are satisfied by F.

The conditions 1), 2), 3) above imply (8.4). To see this, let $Y_1 = [y_{11}, y_{12}], \ldots, Y_m = [y_{m1}, y_{m2}]$ or in abbreviated form $Y = [y_1, y_2]$. Assume $w(F_j(X, Y)) \leqq K_F \max \{w(X), w(Y)\}$. Then

$w(F_j(x,y)) \leqq K_F\, w(Y)$ for real $x \in [x_o, a]$. Since f_j is by definition the real restriction of F_j, we have $f_j(x,y) \in F_j(x,Y)$ whenever $y \in Y$. Therefore $f_j(x,y_1) - f_j(x,y_2) \in F_j(x,Y) - F_j(x,Y)$. Now

$$[a,b] - [a,b] = [-1,1]\, w([a,b]),$$

so

$$|f_j(x,y_1) - f_j(x,y_2)| \leqq w(F_j(x,Y)) \leqq K_F |y_2 - y_1|$$

and therefore

$$|f(x,y_1) - f(x,y_2)| = \max_j |f_j(x,y_1) - f_j(x,y_2)| \leqq K_F |y_2 - y_1| .$$

We notice incidentally that K_F serves as a Lipschitz constant for f.

We conclude that conditions 1), 2), 3) guarantee the existence and uniqueness in D_f^* of a solution to (8.1), (8.2) when f_j is the real restriction of F_j.

If $y_{jo} \in Y_{jo}$ for Y_{jo} properly contained in B_j then the equation

$$Y_{jo} + (X_j^* - x_o)F_j(B) = B_j$$

has a solution X_j^* with $w(X_j^*) > 0$ for each $j = 1, 2, \ldots, m$. By $F_j(B)$ we mean, of course, $F_j(B) = F_j([x_o, a], B_1, \ldots, B_m)$. Define $X^* = [x_o, a] \cap X_1^* \cap, \ldots, \cap X_m^*$.

In this way we can compute an interval, namely X^*, in which existence and uniqueness of a solution y to (8.3), (8.2) is guaranteed.

The First Order Method

Let n be a positive integer and define $X_i^{(n)}, y_{ji}^{(n)}, b_{ji}^{(n)}$ by $y_{jo}^{(n)} = y_{jo}$ and for $i = 1, 2, \ldots, n$

$$\text{(8.5)} \qquad \begin{cases} X_i^{(n)} = \left[x_{i-1}^{(n)},\ x_i^{(n)}\right] = x_o + [i-1, i]\,\dfrac{w(X^*)}{n} \\ b_{ji}^{(n)} = y_{j,i-1}^{(n)} + [0,1]\,\dfrac{w(X^*)}{n} F_j(B) \\ y_{ji}^{(n)} = y_{j,i-1}^{(n)} + \dfrac{w(X^*)}{n} F_j\left(X_i^{(n)},\ b_{1i}^{(n)}, \ldots, b_{mi}^{(n)}\right) \end{cases}$$

In vector notation, dropping the superscript (n), writing h for $\dfrac{w(X^*)}{n}$, and writing $S = [0,1]\, hF(B)$ we simplify the writing of

(8.5) to

$$X_i = x_o + [i-1, i]\,h$$

$$b_i = y_{i-1} + S$$

$$y_i = y_{i-1} + hF(X_i, b_i)$$

so that for $i = 1, 2, \ldots, n$ we have

$$y_i = y_{i-1} + hF(X_i, y_{i-1} + S)$$

This recursion formula expresses in its simplest form, our first order interval method for ordinary differential equations.

The solution y to (8.3), (8.2) clearly satisfies $y(x) \in y(x_{i-1}) + S$ for $x \in X_i$, (that is for each $j = 1, 2, \ldots, m$, $y_j(x) \in y_j(x_{i-1}) + S_j$, etc.) and if $y(x_{i-1}) \in y_{i-1}$, then

$$y(x) = y(x_{i-1}) + \int_{x_{i-1}}^{x} f(x', y(x'))\,dx' \qquad (x \in X_i)$$

so $y(x) \in y_{i-1} + (x - x_{i-1})\,F(X_i, y_{i-1} + S)$ whenever $x \in X_i$. Furthermore, writing $w(y_i) = \max w(y_{ji})$, etc., we find that

$$w(y_i) \leqq w(y_{i-1}) + h\,K_F \max\{h,\ w(y_{i-1}) + c\,h\}$$

where $c = w([0,1]F(B))$, therefore

$$w(y_i) \leqq (\max(c, 1))(e^{K_F(x_i - x_o)} - 1)h \quad . \tag{8.6}$$

Replacing the superscripts, (n), we define for $n = 1, 2, \ldots$ the functions $y^{(n)}$ for all $x \in X^*$, using the fact that

$$X^* = \bigcup_{i=1}^{n} X_i^{(n)} \quad ,$$

by setting

$$y^{(n)}(x) = \begin{cases} y_{i-1}^{(n)} + (x - x_{i-1}^{(n)})F(X_i, y_{i-1}^{(n)} + (X_i^{(n)} - x_{i-1}^{(n)})F(B)) \\ \text{for } x \in X_i^{(n)} \end{cases} \tag{8.7}$$

The functions $y^{(n)}(x)$ are well defined since at $x_i^{(n)}$, the common end point of $X_i^{(n)}$ and $X_{i+1}^{(n)}$, we have

$$y_{i-1}^{(n)} + (x_i^{(n)} - x_{i-1}^{(n)})\, F(X_i, y_{i-1}^{(n)} + (X_i^{(n)} - x_{i-1}^{(n)})F(B)) = y_i^{(n)} .$$

In fact, the functions defined by (8.7) are obviously continuous interval valued functions and are piecewise linear in x, that is, for $0 \leqq t \leqq 1$ we can write

$$y^{(n)}((1-t)\, x_{i-1} + t x_i) = (1-t)\, y_{i-1}^{(n)} + t y_i^{(n)}$$

(However, since $y_{i-1}^{(n)}$ is an interval we do <u>not</u> have $(1-t)y_{i-1}^{(n)} = y_{i-1}^{(n)} - t y_{i-1}^{(n)}$, for $t > 0$.)

We have shown that the interval valued functions $y_j^{(n)}(x)$ defined by (8.5) and (8.7) contain the corresponding components of the solution to (8.1), (8.2); that is, for $n = 1, 2, \ldots,$; and for $j = 1, 2, \ldots, m$, we have, recalling (8.6),

<u>Theorem 5.</u> $\quad y_j(x) \in y_j^{(n)}(x)$ for $x \in X^*$,

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

$$(8.8) \qquad \max_{j = 1, 2, \ldots, m} w\left(y_j^{(n)}(x)\right) \leqq \frac{K}{n} .$$

The following example will serve to illustrate both the geometric and the computational significance of the above result.

Consider the equation

$$\frac{dy}{dx} = y^2$$

and the initial condition

$$y(0) = 1 .$$

The rational interval function defined by $G(Y) = Y^2$ has real restriction $G([y, y]) = f(y) = y^2$. In order to use the same notation as developed for the general case, we define $F(X, Y) = G(Y)$ so

$D_F = \mathcal{I}_{[0,a]} \circledast \mathcal{I}_B$ and F restricted to $B = [0, a] \circledast B_1$ is f. We assume B_1 is an interval of positive width containing the initial value $y(0) = 1$ in its interior and that $a > 0$. The function F clearly satisfies conditions 1), 2), and 3).

Now $F(B) = F([0, a], B_1) = B_1^2$, so call X_1^* the solution of

$$1 + (X_1^*)B_1^2 = B_1 .$$

Since $y(0) \in B_1$, we will always have $0 \in X_1^*$. Set $X^* = [0, a] \cap X_1^*$. Then we can be sure of the existence and uniqueness of a solution for $x \in X^*$. Figure 2 illustrates the geometric significance of the process for determining X^*.

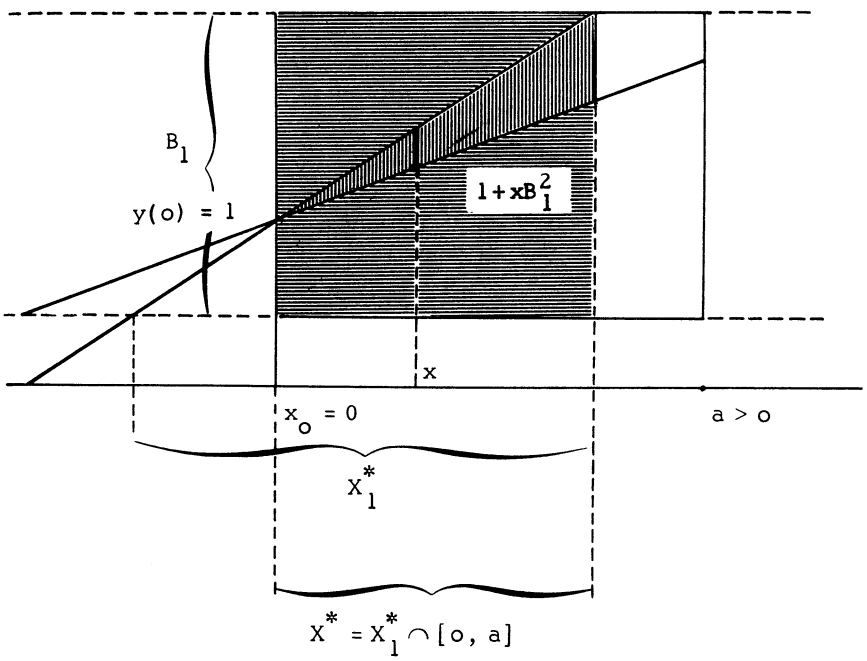

Figure 2

The shaded rectangle is $X^* \otimes B_1$. Let $B_1 = [1/3, 2]$ and $a = 1$, for example; then the lines $y = 1 + 1/9\,x$ and $y = 1 + 4x$ bound a wedge in $X^* \otimes B_1$ containing the solution $y(x)$ of $y' = y^2$, $y(0) = 1$; that is, for $x \in X^* = [0, 1/4]$ we have $y(x) \in 1 + xB_1^2 = 1 + x[1/9, 4]$, since in this case we have

$$1 + X_1^*[1/9, 4] = [1/3, 2]$$

$$X_1^*[1/9, 4] = [-2/3, 1]$$

or if $X^* = [c, d]$, then since $0 \in X^*$ we have $c \leqq o \leqq d$ and

$$[c, d][1/9, 4] = [4c, 4d]$$

so we find $4c = -2/3$, $4d = 1$ or $X_1^* = [-1/6, 1/4]$ and

$$X^* = [0, a] \cap X_1^* = [0, 1] \cap [-1/6, 1/4] = [0, 1/4]$$

Next, we determine for this example, the functions $y^{(n)}$ defined by (8.5) and (8.7). We find that $w(X^*) = 1/4$ so $h = \frac{1}{4n}$ and using $B_1 = [1/3, 2]$, $a = 1$ we determine that

$$F(B) = B_1^2 = [1/9, 4]$$

$$S = [0, 1]\,hB_1^2 = [0, 1]\frac{1}{n}$$

$$X_i^{(n)} = [i-1, i]\frac{1}{4n} = [x_{i-1}^{(n)}, x_i^{(n)}]$$

$$b_i^{(n)} = y_{i-1}^{(n)} + [0, 1]\frac{1}{n}$$

$$y_i^{(n)} = y^{(n)} + \frac{1}{4n} b_i^2 \quad ,$$

therefore

$$y_i^{(n)} = y_{i-1}^{(n)} + \frac{1}{4n}\left(y_{i-1}^{(n)} + [0, 1]\frac{1}{n}\right)^2 \tag{8.9}$$

for $i = 1, 2, \ldots, n$ with $y_o^{(n)} = 1$.

The intervals $y_i^{(n)}$, $i = 1, 2, \ldots, n$, can be computed using (8.9) to obtain the functions $y^{(n)}(x)$. According to (8.7), we have

$$y^{(n)}(x) = \begin{cases} y_{i-1}^{(n)} + \left(x - \frac{(i-1)}{4n}\right)\left(y_{i-1}^{(n)} + [0, 1]\frac{1}{n}\right)^2 \\ \text{for } x \in [i-1, 1]\frac{1}{4n} \end{cases} \tag{8.10}$$

Evaluating (8.6), we find that

$$c = w([0,1]\,F(B)) = w[(0,1][1/9,\ 4]$$

$$= w([0,\ 4]) = 4\ ,$$

and if $Y = [y_1, y_2] \subset B_1 = [1/3, 2]$, then

$$w(F(X,Y)) = w(Y^2) = w([y_1, y_2]^2)$$

$$= w([y_1{}^2, y_2{}^2]) = (y_2 - y_1)(y_2 + y_1)$$

$$= (y_1 + y_2)\, w(Y)\ ,$$

and we can take $K_F = 4$ in order to obtain $w(F(X,Y)) \leqq K_F\, w(Y)$ for $(X, Y) \in D_F = \mathfrak{I}_{[0,1]} \times \mathfrak{I}_{[1/3,2]}$. Making these substitutions in (8.6), we obtain

$$w\left(y_i^{(n)}\right) \leqq \frac{1}{n}\, e^{\frac{i}{n}} \qquad (i = 1, 2, \ldots, n)\ . \tag{8.11}$$

Figure 3 illustrates the construction of the functions $y^{(n)}(x)$ geometrically.

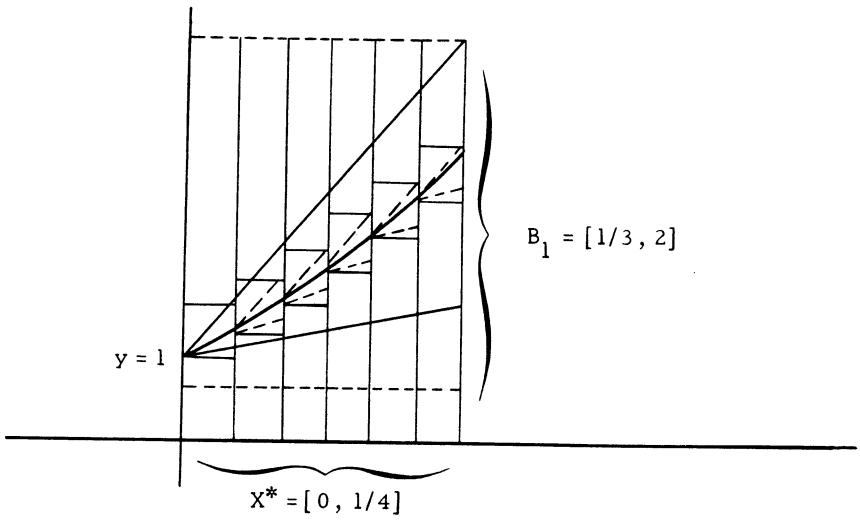

Figure 3

The small rectangles are the b_i for $i = 1, 2, \ldots, n$, while the small dotted triangles are translates of $y_o + (x - x_o)F(B)$ to the intervals y_{i-1}; that is, they represent the interval valued functions of $x \in X_i$ bounding the solutions y to $\frac{dy}{dx} = f(x, y)$ which pass through x_{i-1}, y_{i-1}. If $y(x_{i-1}) \in y_{i-1}$, then $y(x) \in y_{i-1} + (x - x_{i-1})F(B)$ for $x \in X_i$.

The choice of the interval B_1 is seen to affect the width of X^* and the numbers c and K_F in (8.6). The bound expressed by (8.6) on the size of $w(y_i^{(n)})$ was derived in order to prove the convergence of the functions given by (8.7) to the solution of the differential equation. On the other hand, we only need the function F in order to compute the $y_{i-1}^{(n)}$ which determine (8.7) and we will automatically have $y(x) \in y^{(n)}(x)$; so that the interval valued function $y^{(n)}(x)$ gives upper and lower bounds to the solution $y(x)$ at each $x \in X^*$.

For example, setting $n = 10$ in (8.9), we find by interval arithmetic computation that $y_{10}^{(10)} = [1.321\ldots, 1.399\ldots]$ so $w(y_{10}^{(10)}) = .078\ldots$ which is about 1/4 as big as 1/10 e, (compare (8.11)). The exact solution to $\frac{dy}{dx} = y^2$ with $y(0) = 1$, is, of course, given by $y(x) = \frac{1}{1-x}$ and from (8.10), setting $i = 11$, we find that $y^{(10)}(1/4) = y_{10}^{(10)} = [1.321\ldots, 1.399\ldots]$. Thus, $y(1/4) = 4/3 = (1.33\ldots) \in y^{(10)}(1/4)$ - as promised.

Now having computed $y_i^{(n)}$ for $i = 1, 2, \ldots, n$ we can choose $y_n^{(n)}$ as a new initial condition at $x = x_n$, the right hand end point of the interval X^*. Select a new interval B_1 containing $y_n^{(n)}$ in its interior and a new real number a or perhaps use the same $a - x_o$ as before and set $(B_1)_{new} = y_n^{(n)} - y_o + (B_1)_{old}$. If the rectangle so determined still lies in the domain of F, we can proceed as before. In this way we can construct continuations of the functions $y^{(n)}(x)$. To illustrate, we will extend the $y^{(n)}(x)$ obtained for the example we have just treated above. See Figure 4.

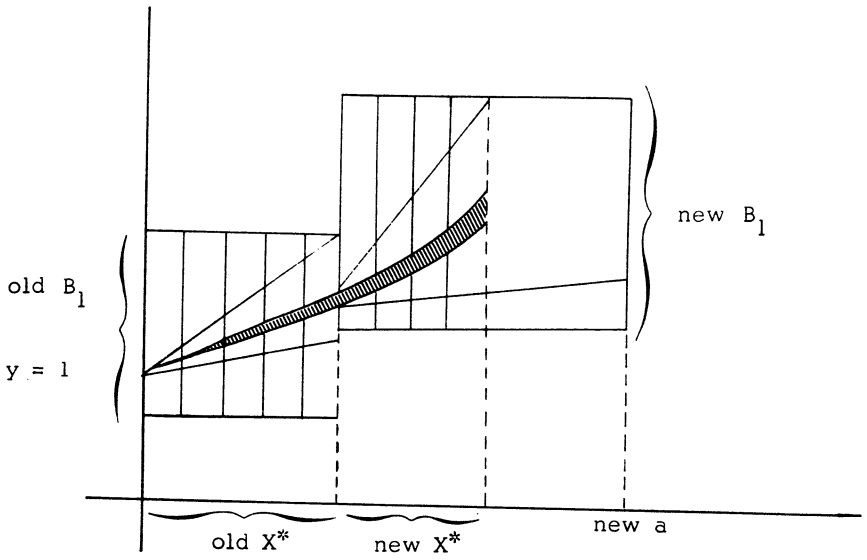

Figure 4

If we choose the new $B_1 = [1, d]$, with $d > y^{(n)}(1/4)$, then $F([1/4, a], B_1) = B_1^2 = [1, d^2]$ and the new X^* is determined by

$$y^{(n)}(\tfrac{1}{4}) + (X_1^* - \tfrac{1}{4})[1, d^2] = [1, d],$$

and

$$X^* = X_1^* \cap [\tfrac{1}{4}, a] = [\tfrac{1}{4}, a^*] .$$

If $y^{(n)}(\tfrac{1}{4}) = [y_1, y_2]$, then a^* is found from

$$y_2 + (a^* - \tfrac{1}{4})\, d^2 = d, \quad d > y_2$$

or

$$a^* = \frac{d - y_2}{d^2} + \frac{1}{4} .$$

Clearly $d = 2y_2$ maximizes a^*, in fact

$$a^* \leqq \frac{1}{4} + \frac{1}{4y_2} \quad .$$

Since $y_2 > \frac{4}{3}$, then $a^* < \frac{7}{16}$. We can again choose a positive integer n and compute the intervals $y_i^{(n)}$ bounding the solution over the new X^*. In fact, for any $0 < \delta < 1$ we can find a finite monotonic sequence of consecutive a^*'s such that the last one is in the interval $[1 - \delta, 1]$ by a finite repetition of the extension procedure. Then the constructed bounding functions $y^{(n)}(x)$ will converge uniformly on $[0, 1 - \delta]$ with increasing n to the solution $y(x) = \frac{1}{1-x}$ and for each n and each $x \in [0, 1 - \delta]$ we will have $y(x) \in y^{(n)}(x)$.

The method we have been discussing is a first order method in the sense of (8.8), that is, the widths of the intervals $y^{(n)}(x)$ for fixed x are $0(n^{-1})$. It should be clear that in the example above, applying the method to the equation $y' = y^2$, the intervals $y^{(n)}(x)$ did not satisfy $w(y^{(n)}(x)) = o(n^{-1})$, for fixed $x > 0$. We will now turn to the investigation of a class of methods such that for each positive integer k, there is a k^{th} order method for constructing interval functions $y^{(k,n)}(x)$ of the real variable x (for x in an interval X^*) which are related to a solution y of (8.1), (8.2) by

$$y(x) \in y^{(k,n)}(x) \quad \text{for} \quad x \in X^*$$

and such that $w(y^{(k,n)}(x)) = 0(n^{-k})$; in fact, for each k there will be a positive real number K_k such that for all $x \in X^*$ and for all positive integers n

$$w(y^{(k,n)}(x)) \leqq K_k \left(\frac{1}{n}\right)^k \quad .$$

The Methods of Order $k > 1$

We will derive a k^{th} order interval method based on local expansion in Taylor series in a fashion similar to the development of $I_{n,k}$ in Section 7.

Let $F = (F_1, \ldots, F_m)$ satisfy conditions 1), 2), 3) stated in the first part of this section. Furthermore, let $F_j^{(\ell)}$, $j = 1, 2, \ldots, m$; $\ell = 0, 1, \ldots, k-1$ be interval valued functions also defined on D_F with real valued real restrictions $f_j^{(\ell)}$ such that $f_j^{(\ell)} = \frac{d}{dx} f_j^{(\ell - 1)}$ on B, that is,

$$f_j^{(\ell)} = \frac{\partial f_j^{(\ell-1)}}{\partial x} + \sum_{r=1}^{m} \frac{\partial f_j^{(\ell-1)}}{\partial y_r} f_r, \quad (\ell = 1, 2, \ldots, k-1),$$

with $f_j^{(0)} = f_j$.

Assume $K_F^{(\ell)}$, $\ell = 0, 1, \ldots, k-1$ are positive real numbers such that $F_j^{(\ell)}$ satisfies conditions 2), 3) with

$$(8.12) \quad w(F_j^{(\ell)}(X, Y_1, \ldots, Y_m)) \leqq K_F^{(\ell)} \max (w(X), w(Y_1), \ldots, w(Y_m)).$$

For example, if $F_j^{(\ell)}$, $j = 1, 2, \ldots, m$; $\ell = 0, 1, \ldots, k-1$ are rational interval functions on D_F then all these conditions are satisfied if $F_j^{(\ell)}$ has real restriction $f_j^{(\ell)}$, a real rational function on B, satisfying $f_j^{(\ell)} = \frac{d}{dx} f_j^{(\ell-1)}$.

Using vector notation again, we put $F^{(\ell)} = (F_1^{(\ell)}, \ldots, F_m^{(\ell)})$, $f^{(\ell)} = f_1^{(\ell)}, \ldots, f_m^{(\ell)})$, etc.

We define the function $A(X, x, Y) = A_1(X, x, Y), \ldots, A_m(X, x, Y))$ by $\mathcal{I}_{[x_o, a]} \circledast D_F$ with $Y = (Y_1, \ldots, Y_m)$ by

$$(8.13) \quad A(X, x, Y) = Y + \sum_{\ell=1}^{k-1} \frac{F^{(\ell-1)}(x, Y)}{\ell!} (X - x)^{\ell} + \frac{F^{(k-1)}(B)}{k!} (X - x)^{k}.$$

The function A will play a role similar to that of $1 + xB_1^2$ in Figure 2 and the dotted triangles of Figure 3.

Recall that $D_F = \mathcal{I}_{[x, a]} \circledast \mathcal{I}_{B_1}, \ldots, \circledast \mathcal{I}_{B_m}$ with y_{jo} in the interior of B_j, $j = 1, \ldots, m$. We choose an X^* such that $w(X^*) > 0$ and $X^* \subset [x_o, a]$ and $A_j(X^*, x_o, y_o) \subset B_j$; $(j = 1, 2, \ldots, m)$.

For every positive integer n, define

$$h_n = \frac{w(X^*)}{n}$$

$$X^{(n)} = x_o + [i-1, i]h_n = [x_{i-1}^{(n)}, x_i^{(n)}], \qquad i = 1, 2, \ldots, n$$

so that $x_o^{(n)} = x_o$ and $x_o^{(n)} = x_o + w(X^*)$ for all n.

Define $y_o^{(n)} = y_o$ for all n and for $i = 1, 2, \dots, n$, define $y_i^{(n)}(x)$ for $x \in X_i^{(n)}$ by the equation

$$(8.14)\quad y_i^{(n)}(x) = A(x, x_o, y_o) \cap \left\{ y_{i-1}^{(n)} + \sum_{\ell=1}^{k-1} \frac{F^{(\ell-1)}(x_{i-1}^{(n)}, y_{i-1}^{(n)})}{\ell!} (x - x_{i-1}^{(n)})^{\ell} + \frac{F^{(k-1)}(X_i^{(n)}, A(X_i^{(n)}, x_{i-1}^{(n)}, y_{i-1}^{(n)}))}{k!} (x - x_{i-1}^{(n)})^{k} \right\}$$

then

$$y_i^{(n)}(x_{i-1}^{(n)}) = y_{i-1}^{(n)} \cap A(x_{i-1}, x_o, y_o),$$

so call $y_i^{(n)}(x_i^{(n)}) = y_i^{(n)}$ for $i = 1, 2, \dots, n$ and (8.14) defines, by finite induction on i, a function $y^{(n)}$ on $x \in X^*$ for each n by setting $y^{(n)}(x) = y_i^{(n)}(x)$ for $x \in X_i^{(n)}$.

The quantities $y_i^{(n)}$, $i = 1, \dots, n$ are each determined by a finite number of evaluations of the $F^{(\ell)}$; substituting $x_i^{(n)}$ for x on the right hand side of (8.14), the left hand side becomes $y_i^{(n)}(x_i^{(n)}) = y_i^{(n)}$.

Recall that we are using vector notation, so that the quantity $y^{(n)}$, for example, is the m-tuple of intervals $(y_{1i}^{(n)}, \dots, y_{mi}^{(n)})$. And $w(y_i^{(n)}) = \max_j w(y_{ji}^{(n)})$, etc.

The formula (8.14) yields a k^{th} order interval method; more precisely, we have the following.

Theorem 6.

The equation

$$y(x) = \bigcap_{n=1}^{\infty} y^{(n)}(x) \qquad x \in X^*$$

defines a function $y \in C^k(X^*)$ satisfying (8.1), (8.2). Furthermore, there is a positive real number M such that for all positive integers n and for all $x \in X^*$,

$$w(y^{(n)}(x)) \leqq M\left(\frac{w(X^*)}{n}\right)^k . \tag{8.15}$$

Proof: From (8.12), (8.13), (8.14) one derives an inequality of the form $w(y_i^{(n)}) \leqq (1 + h_n K)\, w(y_{i-1}^{(n)}) + c h_n^{k+1}$. Then (8.15) follows easily.

In the present situation $w(y_o^{(n)}) = 0$. More generally, (8.15) holds provided $w(y_o^{(n)}) \leqq N(\frac{1}{n})^k$ for some $N > 0$. This fact will permit the continuation of the functions $y^{(n)}$ to a new X^* without losing the inequality (8.15) on the union of the new X^* and the old X^*.

In order to prove the first part of the theorem, we proceed as follows. First, we wish to show that for every $x \in X^*$ and for each pair of positive integers n_1, n_2, the interval valued functions $y^{(n_1)}, y^{(n_2)}$ defined above have non-empty intersection at x, i.e., $y^{(n_1)}(x) \cap y^{(n_2)}(x)$ is non-empty.

In fact, we will use 8.15 to show that for some positive integer N_o and for every $x \in X^*$, $n > N_o$ implies

$$y^{(n\, n_1\, n_2)}(x) \subset y^{(n_1)}(x) \cap y^{(n_2)}(x) \tag{8.16}$$

It is sufficient in order to demonstrate (8.16) to show that for any n_1 there is a large enough N_o such that $n' > N_o$ implies $y^{(n' n_1)}(x) \subset y^{(n_1)}(x)$ for all $n \in X^*$.

From the definition of $y^{(n)}(x)$ it is clear that $y^{(n)}(x) \subset A(x, x_o, y_o)$ for $x \in X^*$. Furthermore, $y^{(n)}(x)$ is non-empty for $x \in X_i^{(n)}$. Since

$$A(x, x_{i-1}^{(n)}, y_{i-1}^{(n)}) = y_{i-1} + \sum_{\ell=1}^{k=1} \frac{F^{(\ell-1)}(x_{i-1}^{(n)}, y_{i-1}^{(n)})}{\ell!} (x - x_{i-1}^{(n)})^k$$

$$+ \frac{F^{(k-1)}(B)}{k!} (x - x_{i-1}^{(n)})^k$$

and

$$(X_i^{(n)}, A(X_1^{(n)}, x_o, y_o)) \subset B$$

it follows that the expression in brackets for $i = 2$ on the right hand side of (8.14), namely

$$y_1^{(n)} + \sum_{\ell=1}^{k=1} \frac{F^{(\ell-1)}(x_1^{(n)}, y_1^{(n)})}{\ell!}(x - x_1^{(n)})^{\ell} + \frac{F^{(k-1)}(X_2^{(n)}, A(X_2^{(n)}, x_1^{(n)}, y_1^{(n)}))}{k!}(x-x_1^{(n)})^{k}$$

is contained in $A(x, x_1^{(n)}, y_1^{(n)})$ for $x \in X_2^{(n)}$.

We claim that for n sufficiently large,

$$A(x, x_{i-1}^{(n)}, y_{i-1}^{(n)}) \subset A(x, x_o, y_o)$$

for $x \in X_i^{(n)}$ and therefore $y^{(n)}(x)$ is non-empty in $X_1^{(n)} \cup X_2^{(n)} \ldots \cup X_i^{(n)}$. Proceeding in this way for $i = 2, 3, \ldots, n$, we finally have $y^{(n)}(x)$ is non-empty in X^*.

Imitating the above argument with $y^{(n_1)}(x) \cap y^{(n'n_1)}(x)$ in place of $A(x, x_o, y_o) \cap \{\ldots\}$ on the right hand side of (8.14), we have for sufficiently large N_o, that $n' > N_o$ implies

$$y^{(n'n_1)}(x) \subset y^{(n_1)}(x) \quad \text{for all} \quad x \in X^* .$$

Thus for each $x \in X^*$, and every finite collection of positive integers $n_1, n_2, \ldots, n_p$, we can conclude that

$$(x, \bigcap_{q=1}^{p} y^{(n_q)}(x))$$

is non-empty and contained in B. By the finite intersection property of the compact set B this means we can define a function y on X^* by

$$y(x) = \bigcap_{n=1}^{\infty} y^{(n)}(x) .$$

For each $x \in X^*$, $y(x)$ is non-empty, in fact it is clearly a real m-tuple. By (8.15),

$$\max_{j=1, 2, \ldots, m} w(y_j(x)) = w(y(x)) \leqq M\left(\frac{w(X^*)}{n}\right)^{k}$$

for all n, hence $w(y(x)) = 0$ for all $x \in X^*$; that is, $y_j(x) = [y_j(x), y_j(x)]$ or $y_j(x)$ is a real number. Notice that $y(x) \in y^{(n)}(x)$ for all n and all $x \in X^*$.

We claim that y satisfies (8.3) with $y(x_o) = y_o$, that is, $y(x) = (y_1(x), \ldots, y_m(x))$ with $\{y_j\}$ satisfying (8.1) and (8.2).

For large enough n, we have for $i = 1, 2, \ldots, n$

$$(8.17) \qquad y_i^{(n)} = y_{i-1}^{(n)} + \sum_{\ell=1}^{k-1} \frac{F^{(\ell-1)}(x_{i-1}^{(n)}, y_{i-1}^{(n)})}{\ell!} h_n^{\ell} + \frac{F^{(k-1)}(X_i^{(m)}, A(X_i^{(n)}, x_{i-1}^{(n)}, y_{i-1}^{(n)}))}{k!} h_n^k$$

where $h_n = \frac{w(X^*)}{n}$ as before.

Recall that $x_{i-1}^{(n)} = x_o + (i-1)h_n$. Now fix $x_{i-1}^{(n)}$, then $x_i^{(n)} = x_{i-1}^{(n)} + h_n$ and let n be large enough so that (8.17) holds, then

$$(8.18) \qquad \left| \frac{y(x_{i-1}^{(n)} + h_n) - y(x_{i-1}^{(n)})}{h_n} - f(x^{(n)}, y(x_{i-1}^{(n)})) \right|$$

$$\leq \frac{1}{h_n} \{ w(y_{i-1}^{(n)} + h_n F^{(0)}(x_{i-1}^{(n)}, y_{i-1}^{(n)})) + w(y_i^n) + \left| h_n^2 \frac{F^{(1)}}{2!} + \ldots + h_n^k \frac{F^{(k-1)}}{k!} \right| \} .$$

Since $k > 1$, then by (8.15) and (8.18) we have

$$\frac{dy}{dx}(x_{i-1}^{(n)}) = f(x_{i-1}^{(m)}, y(x_{i-1}^{(n)})) .$$

In fact, by refinement of the above argument it follows that for all $x \in X^*$, $\frac{dy}{dx}(x) = f(x, y(x))$. Finally, $y \in C^k(X^*)$, since the differentiability of y follows from that of f. This completes the proof of Theorem 6.

9. The Recursive Generation of Taylor Coefficients

In this section, a technique for the recursive generation of Taylor coefficients of arbitrary order for functions defined by ordinary differential systems is presented. The method is applicable to a

wide class of problems and can be programmed for a computer in such a way that only the differential system need be given and the computer will derive the necessary recursion formulas in the form of a subroutine whose execution will provide numerical values of Taylor coefficients of desired order at any desired point. The time required to get the N^{th} coefficient increases at most linearly with N.

The possibility of solving differential equations by power series expansions has been known since B. Taylor (1685-1731). As a practical computational scheme this approach often has been deemed infeasible because of the task of deriving expressions for higher derivations.

On the other hand, good numerical results obtained by local Taylor expansions have been reported for certain differential equations, e.g. [9], [26], [29]. H. T. Davis [5] recommends the use of local expansions to a small fixed number of terms and with a fixed step length (his method of "continuous analytic continuation") particularly for the study of solutions in the neighborhood of singularities.

Recently, a number of programs have been written for the formal (or "algebraic" or "analytic") differentiation of expressions by computers [11], [12]. The input for such programs consists of expressions defining functions and the output consists of other expressions defining the derivatives of the given functions. Even so, the derived expressions for successively higher derivatives usually grow rapidly in length and their direct evaluation may require an exponentially increasing number of operations.

From the point of view of a prescription for computing values of successive derivatives a recursion formula will serve as well as a set of explicit formulas one for each derivative. In fact, it is much more efficient in computing time to save values of lower derivatives and use a recursion formula than to start the computation of each successive derivative from scratch. It has been observed, [9], [26], [29], that recursion formulas for Taylor coefficients can be readily derived for certain types of differential equations.

Actually, the computer can be made to derive the recursion formulas for an extremely wide class of differential systems. In fact, a program has been coded for the IBM 7094 computer by H. R. Jaschke [24] which will accept a differential system as input and produce as output a subroutine. The execution of this subroutine will then produce numerical values of Taylor coefficients up to some desired order at a desired point. After the first N-1 coefficients have been computed, the time required to get the N^{th} coefficient is no more than N times required to get the first coefficient. The subroutine produced by the computer is essentially the machine coding for the recursion formulas and can be used to evaluate the coefficients at any point. Thus a particular solution to a differential system can be computed by successive series expansions (or "analytic continuation") up to terms

of desired order.

The program just described has been used as part of another computer program called "DIFEQ" [24], which carries out the solution of a given differential system (with assigned initial conditions) in rounded interval arithmetic, bounding the remainder term in the truncated Taylor series expansions over intervals it constructs about each new solution point using the methods described in the previous section of this paper. Rounded interval arithmetic also bounds the error due to rounding so that the program is able to supply guaranteed upper bounds on the actual overall error.

Both heuristic analysis and case studies for a variety of nonlinear differential systems indicated that the most efficient choice of the number of terms in the Taylor series carried depended mainly on the number of equivalent significant decimal digits used in the machine arithmetic employed, see section 10 below. For single precision floating point arithmetic on the IBM 7094 (about eight decimal digits) Taylor series expansions out to nine terms are used in the program, DIFEQ, referred to above, [24].

Actual computational experience with the method to be described has indicated the usefulness of a certain normalization procedure in avoiding large ranges of numbers and undue rounding errors, therefore we will introduce normalized Taylor coefficients of a function, Q, as quantities of the form

$$(9.1) \qquad (Q)_j = (1/j!)(d^jQ/dx^j)(\Delta^j) \qquad (j = 0, 1, 2, \ldots)$$

where Δ is a positive real number (the normalization constant) and Q may be regarded either as a real valued function of the real variable x or as a complex valued function of the complex variable x. Or even, as in the program cited above [24], as an interval valued function of the interval variable x.

We will also use, besides (9.1), the notation for a subscripted variable, Q_r,

$$(9.2) \qquad (Q_r)_j \equiv Q_{r,j}$$

when convenient. Of course, $(Q)_0 = Q$.

If Q_1 and Q_2 are functions of x, then

$$(Q_1 \pm Q_2)_j = Q_{1,j} \pm Q_{2,j} \; ;$$

furthermore, it follows from Leibniz's formula and equation (9.1) that

$$(9.4) \qquad (Q_1 \cdot Q_2)_j = \sum_{i=0}^{j} Q_{1,i} \cdot Q_{2,j-i} \; .$$

Writing

$$Q_1/Q_2 = Q_3$$

we have

$$Q_1 = Q_2 Q_3$$

and applying equation (9.4) we obtain

$$Q_{1,j} = \sum_{i=0}^{j} Q_{2,i} \cdot Q_{3,j-i} = Q_{2,0} Q_{3,j} + \sum_{i=1}^{j} Q_{2,i} \cdot Q_{3,j-i}$$

therefore

$$Q_{3,j} = (1/Q_2)\{Q_{1,j} - \sum_{i=1}^{j} Q_{2,i} \cdot Q_{3,j-i}\}$$

or

$$(9.5) \quad (Q_1/Q_2)_j = (1/Q_2)\{Q_{1,j} - \sum_{i=1}^{j} Q_{2,i} \cdot (Q_1/Q_2)_{j-i}\}$$

Equations (9.3), (9.4), and (9.5) are recursion formulas for the j^{th} normalized Taylor coefficients of the sum, difference, product, and quotient of Q_1, Q_2.

If f is a rational function of y and $dy/dx = f(y)$, then the normalized Taylor coefficients $(y)_j$, $j = 0, 1, 2, \ldots$, can be generated using equations (9.3), (9.4), (9.5) as follows:

We choose a particular order in which to evaluate the finite set of arithmetic operations defining f. And for each arithmetic operation in the resulting list we define the result to be a distinct function; call these functions $T_1, T_2, \ldots, T_q$. We then have a set of equations of the form:

$$P_1 *_1 Q_1 = T_1$$

$$P_2 *_2 Q_2 = T_2$$

$$\vdots$$

$$P_q *_q Q_q = T_q = f(y)$$

where $*_s$ is one of the arithmetic operations $+, -, \cdot, \div$ for each $s = 1, 2, \ldots, q$ and where the P_s, Q_s may be constants, the variable y, or a variable T_m with $m < s$. In particular, P_1, Q_1 are each either a constant or y. Applying the appropriate one of the formulas (9.3), (9.4), (9.5) to each equation of the form

$$P_s *_s Q_s = T_s$$

according to which operation $*_s$ is, we obtain a set of expressions for the quantities $(T_1)_j, (T_2)_j, \ldots, (T_q)_j = (f(y))_j$. Finally, we write the equation

$$(y)_{j+1} = (1/(j+1))(f(y))_j \Delta \tag{9.6}$$

to complete the set of recursion formulas.

For example, if

$$dy/dx = f(y) = 3y + y^2 ,$$

then f can be displayed as:

$$y \cdot y = T_1$$

$$3 \cdot y = T_2$$

$$T_1 + T_2 = T_3 = f(y) ;$$

therefore we obtain the recursion relations

$$\sum_{i=0}^{j} (y)_i \cdot (y)_{j-i} = (T_1)_j$$

$$\sum_{i=0}^{j} (3)_i \cdot (y)_{j-i} = (T_2)_j$$

$$(T_1)_j + (T_2)_j = (T_3)_j = (f(y))_j$$

$$(1/(j+1))(f(y))_j \Delta = (y)_{j+1}$$

Notice that, by identifying constants, a reduction in the complexity of the resulting recursion formulas can be achieved. For

example, if Q_2 is constant, (9.3), (9.4), (9.5) become

$$(Q_1 \pm Q_2)_j = Q_{1,j} \qquad (j = 1, 2, \dots)$$

$$(Q_1 \cdot Q_2)_j = Q_2 \cdot Q_{1,j} \qquad (j = 0, 1, 2, \dots)$$

$$(Q_1/Q_2)_j = (1/Q_2)Q_{1,j} \qquad (j = 0, 1, 2, \dots) \quad .$$

In the above example, we may write the recursion formulas as

$$1) \quad y \cdot y = T_1$$

$$2) \quad 3 \cdot y = T_2$$

$$3) \quad T_1 + T_2 = f(y)$$

and for $j = 1, 2, \dots$

$$4) \quad (1/j)(f(y))_{j-1}\Delta = (y)_j$$

$$5) \quad \sum_{i=0}^{j} (y)_i (y)_{j-i} = (T_1)_j$$

$$6) \quad 3 \cdot (y)_j = (T_2)_j$$

$$7) \quad (T_1)_j + (T_2)_j = (f(y))_j \quad .$$

The numerical computation of a set of values for $(y)_1, (y)_2, \dots, (y)_N$ from a given value of y would be carried out by evaluating in order 1), 2), 3) above and then repeatedly evaluating in order 4), 5), 6), 7) for $j = 1, 2, \dots, N$. The quantities $(y)_j$, $(T_1)_j$, $(T_2)_j$, $(f(y))_j$ are all saved for $j = 0, 1, 2, \dots, N$. Thus the storage requirement is $4N$ cells and there are altogether $N(N+1)/2+3N$ multiplications $(N(N+1))/2$ additions, and N divisions required. Each new $(y)_j$ computed requires $j+4$ multiplications, $j+1$ additions, and one division after the previous $(y)_1, \dots, (y)_{j-1}$ have been obtained.

For an equation of the form

$$dy/dx = f(x, y)$$

in which the dependent variable occurs explicitly, we add the equation

$$dx/dx = 1$$

or

$$(x)_1 = \Delta$$

to the list of recursion formulas.

For a system of first order equations

$$dy_r/dx = f_r(x, y_1, y_2, \ldots, y_M) \qquad (r = 1, 2, \ldots, M)$$

with rational f_r, we construct as before the auxiliary equations for the $T_1, T_2, \ldots, T_{q_r}$ for each $r = 1, 2, \ldots, M$ and the corresponding recursion formulas by substitution in (9.3), (9.4), (9.5). The storage required will be

$$(M + \sum_{r=1}^{M} q_r)N$$

cells and the number of additions and multiplications to obtain $(y_r)_j$ for $r = 1, 2, \ldots, M$; $j = 1, 2, \ldots, N$ will be less than $(N(N+1)/2)N_0$, where N_0 is the number of operations needed to obtain the first derivatives alone, i.e., to evaluate the f_r, $r = 1, 2, \ldots, M$.

For differential equations of higher than first order, the usual reduction to a system of first order equations can be made with the substitutions

$$y_r = dy_{r-1}/dx \ .$$

Suppose the system is "autonomous", i.e.,

$$dy_r/dx = f_r(y_1, y_2, \ldots, y_M) \qquad (r = 1, 2, \ldots, M),$$

with x missing from the f_r, and expansion in Taylor series is desired about the point $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_M)$; and suppose we compute the quantity

$$\max_r \left| f_r(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_M) \right| \ .$$

If this quantity is zero, the solution is constant (assuming the f_r are analytic at $(\bar{y}_1, \ldots, \bar{y}_M)$). Otherwise we can put

$$\Delta = 1/\max_r \left| f_r(\bar{y}_1, \ldots, \bar{y}_M) \right| \ .$$

If the system is not autonomous, add the equation

$$\frac{dy_{M+1}}{dx} = 1$$

and substitute y_{M+1} for x in f_r, $(r = 1, 2, \ldots, M)$, and again we can use the above choice of Δ.

In floating point computation the number 1.0 is symmetrically placed with respect to the range of exponents and the above choice of Δ "normalizes" to 1 the vector of "rates of change" in the solution components at each point where a series expansion is carried out.

The solutions of rational differential systems include virtually all the special functions used in scientific computing. They include $\sin x$, $\cos x$ which satisfy

(9.7)
$$dy_1/dx = y_2$$
$$dy_2/dx = -y_1$$

$\exp x$ which satisfies

(9.8)
$$dy/dx = y$$

x^a which satisfies

(9.9)
$$dy/dx = ay/x$$

etc., etc., etc.

In fact, given a differential equation of the form

(9.10)
$$dy/dx = F(y)$$

with an expression for F containing a finite number of rational operations and compositions of non-rational functions which themselves satisfy rational differential equations, we can substitute new variables for each of the non-rational functions and add their defining rational differential equations to obtain a rational system in place of (9.10).

For example, given

(9.11)
$$dy/dx = \alpha \cos(\exp y^2) + y^{-.15},$$

substitute

(9.12)
$$y_1 = \exp y^2$$
$$y_2 = \cos y_1$$
$$y_3 = y^{-.15}$$

then equation (9.11) becomes

(9.13)
$$dy/dx = \alpha y_2 + y_3$$

and we derive the additional equations needed from the "chain rule" for differentiating composite functions,

(9.14)
$$df(g(x))/dx = f' dg/dx$$

and known rules for the differentiation of the functions exp, cos, etc. Thus

(9.15)
$$dy_1/dx = (\exp y^2)(2y)(dy/dx)$$

or

(9.16)
$$dy_1/dx = (y_1)(2y)(dy/dx)$$

and

$$dy_2/dx = -(\sin y_1)(dy_1/dx)$$
$$d(\sin y_1)/dx = (\cos y_1)(dy_1/dx)$$

or substituting

(9.17)
$$y_4 = \sin y_1$$

we obtain

(9.18)
$$dy_2/dx = -y_4(dy_1/dx)$$

(9.19)
$$dy_4/dx = y_2(dy_1/dx)$$

and finally we have

(9.20)
$$dy_3/dx = -.15 y_3/y .$$

Now equations (9.13), (9.16), (9.18), (9.19), (9.20) are rational in

the variables y, y_1, y_2, y_3, y_4 so that the technique of the previous section applies and will provide recursion formulas for the determination of the Taylor coefficients $(y)_j$ $(j = 0, 1, 2, \ldots)$.

This process of reduction of a differential system to a rational differential system by the substitution of new variables and the addition of rational defining equations can evidently be carried out for any system of equations expressed in a finite number of rational operations and compositions of functions which themselves satisfy reducible systems.

An alternative approach to the handling of non-rational functions occuring in the differential systems is available.

Rather than adding more differential equations to a given system in order to define such functions as sin, cos, exp, etc., we can instead proceed as follows. We prepare interval-valued extensions of these functions in the form of computer "subroutines" as discussed in section 4 above. In addition, we augment the set of basic formulas for the j^{th} normalized derivatives of sums, products, and quotients, (9.3), (9.4), (9.5), by the addition of formulas which recursively define higher derivatives of the non-rational functions in question.

For the functions, sin and cos, for example, we have, using the notation defined by (9.1), the following formulas:

$$(9.21) \qquad (\sin Q)_j = (1/j) \sum_{i=0}^{j-1} (i+1)(\cos Q)_{j-1-i}(Q)_{i+1}$$

$$(9.22) \qquad (\cos Q)_j = (1/j) \sum_{i=0}^{j-1} (i+1)(\sin Q)_{j-1-i}(Q)_{i+1} \quad .$$

In order to obtain $(\sin Q)_{10}$, for example, we would compute successively the pairs of values, using both (9.21) and (9.22), $(\sin Q)_1, (\cos Q)_1, (\sin Q)_2, \ldots, (\sin Q)_9, (\cos Q)_9, (\sin Q)_{10}$.

The formulas (9.21), (9.22) are derived using the chain rule (9.14) and formulas (9.4), (9.6).

10. The Automatic Selection of Approximation Parameter Values

In section 8 above, we presented a procedure with an explicit formula (8.14) for the computation of intervals containing solution values for systems of ordinary differential equations with given initial conditions.

The procedure is based on local expansions in Taylor series

with remainder term of order k. The remainder term is to be evaluated by interval computation over regions denoted by B, where $B = (B_1, B_2, \ldots, B_m)$ is a vector of intervals $B_1, B_2, \ldots, B_m$ for a differential system of order m.

The procedure further requires the determination of an interval X^* such that for $x \in X^*$, the solution $y(x) = (y_1(x), y_2(x), \ldots, y_m(x))$ lies in the region B. That is, $y_1(x) \in B_1, \ldots, y_m(x) \in B_m$.

Finally, the interval X^* is subdivided into n parts and formula (8.14) is evaluated yielding intervals $y_i^{(n)}$ containing $y(x_i^{(n)})$ whose widths are bounded, according to (8.15), by

$$w(y_i^n) \leq M\left(\frac{w(X^*)}{n}\right)^k .$$

In section 9 above, we discussed a procedure for getting the computer to derive recursion formulas in the form of "coded subroutines" which will evaluate the "normalized" Taylor coefficients

$$(Y)_j = 1/j!\, F^{(j-1)}(Y)\Delta^j \tag{10.1}$$

required by (8.14). In (10.1) and throughout this section the quantities Y and $F^{(j-1)}(Y)$ are assumed to be m-dimensional interval vectors $Y = (Y_1, \ldots, Y_m)$, $F = (F_1, \ldots, F_m)$.

Actually the formula (8.14) as given does not exhibit the normalization factor Δ. We can rewrite the sum occurring in (8.14) as

$$Y + \sum_{j=1}^{k-1} (Y)_j \left\{\frac{(x - x_{i-1}^{(n)})}{\Delta}\right\}^j + (Y)_k \left\{\frac{x - x_{i-1}^{(n)}}{\Delta}\right\}^k \tag{10.2}$$

We have omitted the arguments of the functions $F^{(j-1)}$ in writing (10.2) and have used the simpler notation afforded by (10.1). The last term in (10.2) is split off because it is the "remainder" term and has a different set of arguments than the other terms, see (8.14).

As in the previous section we will assume here that the differential system has been made autonomous by adding, if necessary, the equation

$$\frac{dy_{m+1}}{dx} = 1$$

and replacing x by y_{m+1} in the functions $F_1, F_2, \ldots, F_m$ of the system

$$\frac{dy_r}{dx} = F_r(x, y_1, \ldots, y_m), \qquad (r = 1, 2, \ldots, m) .$$

The resulting system to be solved has the vector form

$$\frac{dy}{dx} = F(y), \quad y(x_0) = y_0 . \tag{10.3}$$

We will consider now the design of a program based on (8.14).

We assume the program will use subroutines which perform the required rounded interval arithmetic using normalized floating point machine numbers whose fractional part consists of a fixed number of binary digits, s; (on the IBM 7094 computer, s = 27 for single precision floating point arithmetic and s = 54 for double precision floating point arithmetic).

The approximation parameters intrinsic to the method in question are s, k, n, B, Δ and X^*. Values of these parameters must be selected at the initial point x_0, y_0 and at each subsequent point x_i, Y_i where a Taylor expansion is to be carried out. We will discuss the automatic determination of all these values by the computer during the course of a particular solution.

Providing only that the selection of X^* satisfies the inclusion relation

$$Y(X^*) \subset B , \tag{10.4}$$

any set of values for s, k, n, B, Δ, and X^* yields a computation resulting in intervals Y_i guaranteed to contain the exact solution at x_i

$$y(x_i) \in Y_i .$$

Furthermore, (8.14) will produce for any desired $x \in X^*$, an interval $Y(x)$ containing the exact solution at x

$$y(x) \in Y(x) .$$

The manner in which s, k, n, B, Δ and X^* are chosen will determine the widths of the computed intervals Y_i and the amount of time required for the computation. The storage space required in the computer depends mainly on s and k.

The problem of "overflow" and "underflow", i.e., computations producing numbers exceeding the range of exponents in machine floating point representation, depends mainly on k, B and Δ and also on

how far the solution is carried in case the range of actual solution values or the range of Taylor coefficient values grows large.

The design of most computers, with their fixed word length arithmetic, restricts a sensible choice of s, the number of binary places carried in the arithmetic operations, to values corresponding to single precision, double precision, etc. so we leave that decision to the "user" of our program and we will determine the rest of the parameters as functions of s.

In section 9 above we have already mentioned a choice for Δ, namely

$$\Delta = \Delta_0 = 1/\max_r |F_r(Y_i)| . \tag{10.5}$$

If, during the computation of a set of Taylor coefficients (Y_1), $(Y)_2, \ldots, (Y)_k$, using a particular value of Δ, say Δ_p an overflow or underflow should occur, the program can be designed to try again with an alternative Δ, say $\Delta_{p+1} = 2\Delta_p$ or $\Delta_{p+1} = 1/2\Delta_p$ depending on the situation. The program can be designed to terminate the computation printing a message concerning the source of trouble in case overflow persists, say, after a certain number of such rescalings have been tried.

The interval function A occurring in (8.14) is defined by the expansion (8.13) and has a remainder term with a factor $(Y)_k$ evaluated over the region B, (see also figure 3).

In using a value of n larger than 1, we arrive at a first interval solution value $y_i^{(n)}$ using (8.14) with some predetermined B which is wider than we would get by using $n = 1$ and $B = A(X_1^{(n)}, x_0, y_0)$.

In fact if we could somehow find a B such that

$$y_1^{(1)}(X^*) = B \tag{10.6}$$

for a given $X^* = X_1^{(1)}$, then $y_1^{(1)}$ would be the narrowest possible interval the method could produce at the end of one step for a given k.

The cost of using a new B for each new interval solution value is the added cost of evaluating the Taylor coefficients $(Y)_j$, $j = 1, 2, \ldots, k$, over a new B at each solution point instead of once for each set of n solution points.

For $n > 1$, however, we would need n sets of coefficients, within a given B, evaluated over the successive interval solution points $y_i^{(n)}$, anyway.

All things considered, (or at least several), we will set $n = 1$ and seek to determine, for given k and s, a region B reasonably close to that which minimizes the width of a one step interval solution over X^*, with X^* chosen as wide as possible keeping

the contribution of the remainder term to the "relative" width of the resulting interval solution y_1 about as small as 2^{-s}.

This is similar to a "variable step size" procedure based on keeping the "local truncation error" approximately constant.

For the choice $n = 1$, the intersection in formula (8.14) automatically yields the second term and can be dropped since we are going to choose X^* and B subject to the condition that

$$A(X^*, x_0, y_0) \subset B \,. \tag{10.7}$$

In this case, the formula for the interval solution y_1 at $x_0 + x \cdot \Delta$, for $0 \leq x \leq \frac{w(X^*)}{\Delta}$, and using (10.1), becomes

$$y_1 = y_0 + \sum_{j=1}^{k-1} (Y)_j x^j + (Y)_k x^k \tag{10.8}$$

where the coefficients $(Y)_j$, $j = 1, 2, \ldots, k-1$, are evaluated at the previous interval solution, y_0; the coefficient $(Y)_k$ is evaluated over the region $A(X^*, x_0, y_0)$ given by

$$A(X^*, x_0, y_0) = y_0 + \sum_{j=1}^{k-1} (Y)_j \left(\frac{X^* - x_0}{\Delta}\right)^j + (Y)_k \left(\frac{X^* - x_0}{\Delta}\right)^k \tag{10.9}$$

In (10.9) the coefficients $(Y)_j$, $j = 1, 2, \ldots, k-1$ are the same as in (10.8); however, the coefficient $(Y)_k$ in (10.9) is to be evaluated over the region B. The interval $X^* - x_0$ can also be written $X^* - x_0 = [0, w(X^*)] = w(X^*)[0, 1]$.

After describing our procedure for automating the selection of B and X^* we will finally get to the choice of k.

The remainder term in (10.8) is $(Y)_k x^k$ and has width $w((Y)_k)x^k$, (see the end of section 4 above). Since we are hoping for a wide X^*, i.e., a large "step size", we will compare the width of the remainder term with an estimate on the change in the solution values rather than the solution values themselves. We take this estimate to be $(Y)_1 x$. On account of our normalization, ((10.5), (10.1)) we therefore seek to determine $X^* = [x_0, x_0 + \bar{x} \cdot \Delta]$ with $\bar{x}$ satisfying

$$w((Y)_k)\,\bar{x}^k = 2^{-s}\bar{x} \,, \tag{10.10}$$

where $(Y)_k$ is evaluated over the region $A(X^*, x_0, y_0)$.

For numerical evaluation of the solution we use the "nested" form for the polynomials (10.8) and (10.9) in order to obtain smaller

interval widths! See sections 3 and 4 above, and also compare (7.9).

From (10.9) we can expect the width of A and hence of $(Y)_k$ in (10.8) to increase with the widths of the components of B, therefore we want the narrowest set of intervals for $B = (B_1, B_2, \ldots, B_m)$ satisfying (10.7), namely $B_r = A_r(X^*, x_0, y_0)$. In the subsequent discussion the arguments x_0, y_0 are fixed and we will drop them and write simply $A(X^*) = A(X^*, x_0, y_0)$.

We recapitulate the implicit relations we have derived for the determination of the desired X^* and B.

$$X^* = x_0 + [0, \bar{x} \cdot \Delta] \tag{10.11}$$

$$A(X^*) = y_0 + \sum_{j=1}^{k-1} (Y)_j [0, \bar{x}]^j + (Y)_k(B)\,[0, \bar{x}]^k \tag{10.12}$$

$$w((Y)_k(A(X^*)))\,\bar{x}^k = 2^{-s}\bar{x} \tag{10.13}$$

$$B = A(X^*)\ . \tag{10.14}$$

In any case, if we use any X^*, B which satisfy (10.7), in the evaluation of (10.9) and (10.8) we will obtain an interval solution y_1 which contains the exact solution at the corresponding argument, (Theorem 6). Notice, in (10.12), that for a fixed B which properly contains y_0 we will have $A(X^*) \subset B$ for sufficiently small $\bar{x}$.

There are many possible approaches to an iterative solution of the equations (10.11), (10.12), (10.13), (10.14) for X^* and B. We have tried several with varying degrees of success.

The method we present here is slightly less elaborate than the one actually used in the program "DIFEQ", [24], referred to in section 9, but it has also been used successfully.

For the first solution point computed from the given initial values y_0, we proceed as follows.

Assume that a region B and an X^* will be chosen such that $w((Y)_k(A(X^*)) = 1$ with $A(X^*)$ given by (10.12). Then from (10.13) we compute that

$$\bar{x}^{k-1} = 2^{-s}$$

or

$$\bar{x} = (2^{-s})^{1/k-1}\ . \tag{10.15}$$

Using this value of $\bar{x}$, we set

$$B = y_0 + (Y)_1[0, \bar{x}] \quad . \tag{10.16}$$

We then compute X^*, $A(X^*)$ by (10.11), (10.12) and test the relation (10.7). If it is satisfied, we compute $(Y)_k(A(X^*))$ and evaluate the interval solution y_1 by (10.8) at any desired points $x \in X^*$, in particular, at $x_0 + \bar{x} \cdot \Delta$ for the beginning of a new Taylor expansion.

In case the relation (10.7) is not satisfied, we compute a new $\bar{x}$ using (10.13) and substitute our previously computed $A(X^*)$ for B.

We then compute a new X^*, $A(X^*)$ by (10.11), (10.12) and again test (10.7).

Following satisfaction of (10.7) we always proceed the same way.

In case of failure to satisfy (10.7) we alternate between the procedure just described and the following: Replace $\bar{x}$ by $(1/2)\bar{x}$ and reevaluate X^*, $A(X^*)$ by (10.11), (10.12) again testing (10.7).

For successive solution points, after the first, we use the same process where y_0 is now the last computed interval solution except that instead of (10.15) for the first trial value of $\bar{x}$, we put

$$B = y_0 + (Y)_1[0, x^*] \tag{10.17}$$

where x^* is the final value settled upon for $\bar{x}$ in the computation of the last computed interval solution (i.e., what we now call y_0).

Using the B computed by (10.17) we get a trial value of $\bar{x}$ from

$$w((Y)_k(B))\bar{x}^k = 2^{-s}\bar{x} \quad . \tag{10.18}$$

We the compute X^*, $A(X^*)$ from (10.11), (10.12) and repeat the same process as was described for the first solution point, alternately replacing B by $A(X^*)$ and cutting $\bar{x}$ in half until $A(X^*) \subset B$.

It can be proved that after a finite number of replacements of B by $A(X^*)$ not greater than the order of the differential system, m, the resulting region B will at least have the correct number of dimensions.

A replacement of B by $A(X^*)$ with $A(X^*) \not\subset B$ will usually lead to a smaller $\bar{x}$ and, of course, so will halving $\bar{x}$, hence we can expect that eventually an $\bar{x}$ will be reached which is sufficiently small to cause $A(X^*) \subset B$.

Actually, in practice, we have observed that the process most often terminates after the first replacement of B by $A(X^*)$.

Should an overflow, an underflow, or an attempted division by an interval containing zero occur during the determination of the

Taylor coefficients being evaluated over a trial region B, which would happen for example if B is large enough to contain (or even approach too closely) a singularity, then the program we are designing should be made to try again with a reduced region.

This can be done effectively by halving $\bar{x}$ or x^* and repeating the computation of B. A limit to the number of such halvings allowed for each new solution point should be set no greater than the number of binary places in the arithmetic used. The number of allowed halvings will affect the distance of closest possible approach by the program to a singularity of the solution.

The main difference between the procedure just described for the automatic selection of X^* and B and the one used in the program, "DIFEQ", [24] is that in [24] we defined a certain quantity d by the relation

$$w((Y)_k(A(X^*))) = \bar{x}^d w((Y)_k(B)) \tag{10.19}$$

and in place of (10.13) we use

$$w((Y)_k(B))\,\bar{x}^{k-1+d} = 2^{-s} \tag{10.20}$$

in order to compute $\bar{x}$.

In many cases, this variation gave a suitable X^* and B on the first iteration. The form of the relation (10.19) was motivated by a number of observations on actual computations and is of no particular importance for our purpose here. We have mentioned this variation because we will quote some numerical results obtained using DIFEQ, [24], in the last section of this paper.

We have disposed of the selection (either by legislation or an automatic process to be carried out by the computer) of the approximation parameters, Δ, n, B, X^*, intrinsic to our interval method for solving ordinary differential equations. We have left to describe the selection of k, the number of terms to be used in the Taylor expansions.

The procedures we have described thus far will yield solution intervals whose widths vary little with k except that for very small k, like $k = 1$ or $k = 2$, we can expect a considerable growth of interval widths due to accumulation of roundings by the rounded interval arithmetic over the necessarily large number of steps required. For $k = 1$, we will get very small values of $\bar{x}$ from (10.13).

On the other hand, the time required to carry out the computation in order to reach a given value of the independent variable x will depend very much on k.

As can be seen from section 9 above, the time required to get a set of values of the coefficients $(Y)_1, (Y)_2, \ldots, (Y)_k$ is roughly

proportional to k^2 for a non-linear differential system.

Assuming that the computation time for a given point is proportional to the time spent obtaining sets of values of the Taylor coefficients, then we wish to minimize $T(k) = k^2 N(k)$ where $N(k)$ is the number of points required to reach a given value of x by the successive expansions in interval-valued Taylor series which we have described.

To get an approximate solution of this problem, we make the simplifying assumptions that $w((Y)_k(A(X^*))) = 1$ in (10.13) and that all the steps will satisfy $\bar{x} = 1/N(k)$. Then we have the following relation from (10.13)

$$\left(\frac{1}{N(k)}\right)^{k-1} = 2^{-s} \quad . \tag{10.21}$$

This gives

$$N(k) = 2^{\frac{s}{k-1}} \quad . \tag{10.22}$$

So we wish to minimize

$$T(k) = k^2 \exp\{\frac{s}{k-1} \log_e 2\} \tag{10.23}$$

as a function of k.

We take (10.23) to define a function of a continuous variable k and we put $T'(k) = 0$ so that we determine k from

$$2k - k^2\{\frac{s}{(k-1)^2} \log_e 2\} = 0 \quad . \tag{10.24}$$

By this reasoning, then, we should choose k to be, say, the nearest integer to the solution of (10.24), namely

$$k = \text{nearest integer to } \{2 + (.346\ldots)s\} \quad . \tag{10.25}$$

For example, (10.25) gives

$k = 11$ for $s = 27$, (single precision);

$k = 21$ for $s = 54$, (double precision).

Case studies were made using the program, "DIFEQ", [24], with variable k for a number of differential systems and the results indicated that $k = 9$ gave close to the fastest computation times in

all the equations tested with the single precision rounded interval airthmetic on the IBM 7094 computer. We were able to determine this by testing a clock built into the computer and we printed out the actual computing times for various values of k.

As a result of the tests made we decided to incorporate the fixed choices of $k = 9$ and $k = 19$ in our single and double precision verisons of "DIFEQ", [24], respectively.

More than likely, further research will result in a more sophisticated automatic procedure for the determination by the computer of an efficient choice of k depending on the particular equation being treated and even on the particular point at which the Taylor expansion is being carried out.

11. Numerical Results Obtained with the Interval Differential Equations Program

In section 8 above, we derived a famliy of k^{th} order methods for computing intervals containing values of solutions of ordinary differential equations.

The "interval solutions" are obtained in a step by step fashion by means of expansions at each step in Taylor series with remainder, truncated at the k^{th} term. The remainder term is evaluated in interval arithmetic over regions made up of intervals constructed about each new solution point. The step size is chosen so that the solution remains in the constructed region for all intermediate values between one solution point and the next.

In section 9 above, we presented a means by which the computer can be made to derive recursion relations for the efficient computation of both real and interval values of Taylor coefficients as required by the interval differential equations method.

In order to take into account the finite precision of machine arithmetic, "rounded interval arithmetic", sections 2 and 3 above,is used to evaluate the formulas, (8.14), etc.

We have made several references to an operating machine program, "DIFEQ", [24], which incorporates the procedures just described. In this section, we will quote some numerical results obtained with the program. But first we will summarize the features of the program which result from its incorporation of the methods described in this paper.

1. Along with each computed approximate solution value Y, the program produces a rigorous upper bound, e, on the total error. If y is the exact solution value approximated by Y, then $|y - Y| \leq e$ holds. In order to obtain results in this form, the program, which computes with interval numbers, simply prints the midpoint, Y, and one half the width, e, of the interval solution.

2. The only required input to the program is the differential system to be solved and the initial values defining the particular solution desired. The program during its execution on the computer determines all the approximation parameter values intrinsic to the method such as initial "step size", subsequent "step sizes", etc. If the user desires, he may specify values of the independent variable at which he would like solution values; otherwise the program will print each solution point it obtains along with a set of Taylor coefficients for interpolating intermediate points.

3. In specifying initial conditions and equation constants for a particular solution, inexact data is allowed, i.e., data of the form, $C \pm e$. These "initial errors" will also be taken into account by the program. The program will then compute intervals which contain simultaneously all solutions beginning with real values chosen from the given intervals of initial conditions.

Example 1.

Davis, [5], has computed a table, (his Table I, Appendix 4), of values of solutions of the equation defining the "first Painleve transcendent",

$$d^2y/dx^2 = 6y^2 + \lambda x \tag{11.1}$$

for the initial conditions $x_0 = 0$, $y_0 = 1$, $y_0' = 0$ and for each of the values $\lambda = 0, 1, 2, 3, 4, 5$.

The values are given to five significant decimal digits for $x = 0(.01)1.00$ except near 1.00 where only three or four figures are given.

Using the single precision version of our interval differential equations program, "DIFEQ", values and error bounds were computed for $x = 0(.01)1.00$, $\lambda = 0, 1, 2, 3, 4, 5$.

This computation required a total of 1.5 minutes on the computer.

The largest error bounds we obtained were 2 in the seventh decimal digit for y and 4 in the seventh digit for y'. These occurred at $x = 1.00$, $\lambda = 5$.

By comparison, we were able to verify that the results given in [5], (Table I, Appendix 4), are substantially correct except that a number of values are off in the last place or two. For $\lambda = 0$, for example, the values of y at $x = 0.79$, 0.80 are given in [5] as 5.5570, 5.8277 whereas the correct values were determined by DIFEQ to lie in the ranges $5.558583 \pm 1.6 \cdot 10^{-6}$, $5.829493 \pm 1.7 \cdot 10^{-6}$, respectively.

Example 2.

In [5] on p. 480 there is a table of values to ten places of $y(x) = 2/(1 + e^{x^2})$, for $x = 0(.02)1.00$; this function is the solution to the equation

$$y' = xy(y - 2) \tag{11.2}$$

with the initial condition $y(0) = 1.0$. The equation (11.2) is used in [5] to illustrate various numerical methods for differential equations.

We submitted equation (11.2) with $y(0) = 1.0$ to DIFEQ; and interval solutions were computed at $x = 0(.02)1.00$.

The time required for the computation was 0.1 minutes on the computer.

The maximum width of the computed intervals occurred at $x = 1.00$, where the program obtained

$$y(1) = .53788284 \pm 1.3\ 10^{-7}$$

or

$$y(1) \in [.53788271, .53788297] .$$

The correct value, according to [5], to ten places is

$$y(1) = .5378828427 .$$

Example 3.

The equation

$$y' = y^2 \tag{11.3}$$

was given to the program DIFEQ with $y(0) = 1$.

The program computed interval solutions at 87 values of x; the successive values of x at which solutions were computed and at which Taylor expansions were made by the program grew close together as x approached the value 1.

The solution of (11.3) with $y(0) = 1$ is, of course,

$$y(x) = 1/(1 - x) .$$

The last interval solution value computed was at $x = .99986639$ where the program obtained

$$y = 7486.06 \pm 5.68$$

or

$$y \in [7480.38,\ 7491.74]$$

and printed the message:

PROGRAM UNABLE TO BOUND DERIVATIVES OVER -

X = .99986639

Y = 7486.0641 ± 5.6762085 .

The correct solution value rounded to six figures at $x = .99986639$ is $y = 7484.47$.

Since the method used by the program, sections 8-10 above, includes the bounding of derivatives of the exact solution over the whole interval of values of the independent variable between successive solution points, it cannot integrate past a singularity where values become infinite. It will stop short, in fact, at a point where the range of machine numbers is exceeded, see section 10 above.

In [5], (especially pp. 263-266), Davis discusses the numerical analytic continuation of solutions into the complex plane in order to integrate around singular points.

Example 4.

The differential equations for the so-called restricted problem of three bodies are usually given as

$$x'' - 2y' = x - \frac{\mu(x - 1 + \mu)}{r^3} - \frac{(1 - \mu)(x + \mu)}{R^3} \tag{11.4}$$

$$y'' + 2x' = y - \frac{\mu y}{r^3} - \frac{(1 - \mu)y}{R^3}$$

where $$r = \{(x - 1 + \mu)^2 + y^2\}^{\frac{1}{2}}$$

and $$R = \{(x + \mu)^2 + y^2\}^{\frac{1}{2}}$$

Putting $\mu = .01215$, the equations become a mathematical model for the motion in a plane of a space vehicle in free fall in the earth-moon gravitational system.

With this value of μ and the initial conditions

$$\left.\begin{aligned} x &= 1.2 \\ y &= 0 \\ x' &= 0 \\ y' &= -1.0493575 \end{aligned}\right\} \text{at } t = 0 \quad \text{(indep. var.)}$$

We made a study of the effect of changing the number of terms carried in the Taylor expansions in the single precision version of DIFEQ. The table below shows the results of the study. Since in equations (11.4), one of the dependent variables is denoted by "x", we denote the independent variable by "t". For each of the values 5, 6, 7, 8, 9 for k, the number of terms carried in the Taylor expansions, we have called for DIFEQ to integrate the system (11.4), with the initial conditions given above, up to the value t = 1.00. The table shows the computation time in minutes, the number of intermediate points, (i.e., the number of Taylor expansions which the program actually used in reaching t = 1.0), and the maximum error bound produced by the computation, (which occurred in every case in x' at t = 1.0).

No. of terms in expansion	Computation time (min.)	No. of computed points	Max. error bound
5	4.26	171	$1.65 \cdot 10^{-5}$
6	3.48	98	$1.01 \cdot 10^{-5}$
7	3.21	67	$7.30 \cdot 10^{-6}$
8	3.15	51	$6.18 \cdot 10^{-6}$
9	3.38	43	$5.61 \cdot 10^{-6}$

A similar case study was made with the same equations (11.4) but with the initial conditions $x = -1.98012 \cdot 10^{-2}$, $y = -1.51062 \cdot 10^{-2}$, $x' = 9.5560068$, $y' = -4.856878$ with the following results:

No. of terms in expansion	Computation time (min.)	No. of computed points	Max. error bound (at t = .01)
7	4.28	89	$2.95 \cdot 10^{-5}$
8	4.09	66	$2.46 \cdot 10^{-5}$
9	4.00	52	$2.33 \cdot 10^{-5}$

Example 5.

On pp. 85-86 of [16], Henrici considers the equation $y' = -16xy$ with the initial condition $y(-0.75) = 0.00221\ 59242\ \ldots$ He computes numerical approximations to the solution values at $x = -0.50(.25)0.75$ using the Runge-Kutta method with a constant step $h = 2^{-p}$ for $p = 4(1)9$; and lists actual errors and errors predicted by analytical methods developed in the book. The agreement is very good; for example, the errors for $p = 7$ are:

x = 0		x = 0.25		x = 0.50	
actual	predicted	actual	predicted	actual	predicted
18.10^{-8}	19.10^{-8}	11.10^{-8}	12.10^{-8}	2.10^{-8}	2.10^{-8}

Using the program DIFEQ, [24], we obtained interval solutions with the following actual errors in midpoints and automatically computed error bounds, (half the widths of the interval solutions):

x = 0		x = 0.25		x = 0.50	
actual	bound	actual	bound	actual	bound
2.10^{-8}	14.10^{-8}	1.10^{-8}	23.10^{-8}	$.5.10^{-8}$	105.10^{-8}

The time required to obtain the single precision interval solutions at $x = -0.50(.25)0.75$ was 0.14 minutes on the IBM 7094 computer.

A corresponding run was made with the double precision version of DIFEQ resulting in actual errors in midpoints and automatically computed bounds on these errors as follows:

x = 0		x = 0.25		x = 0.50	
actual	bound	actual	bound	actual	bound
2.10^{-16}(?)	1.10^{-15}	$1.2.10^{-16}$(?)	2.10^{-15}	$.6.10^{-16}$(?)	8.10^{-15}

The "actual errors listed are only estimates based on departure from symmetry. We did not have the exact results available to 16 places.

The time required for the double precision run was 0.36 minutes, on the IBM 7094 computer.

12. Conclusions

If a real number, x, is defined as the result of a finite sequence of arithmetic operations beginning with a finite collection of real numbers with known decimal representations, then the execution by the computer of the corresponding finite sequence or rounded interval arithmetic operations produces an interval, **X**, containing the

real number, x. By carrying enoutgh places, the width of X can be made arbitrarily small. If s binary places are carried, then the width of X will be proportional to 2^{-s}.

Remainder terms in the truncation of Taylor series, etc., can be bounded over regions with interval computations.

In fact, the concept of interval valued functions provides a basis for the design of computational schemes for digital computers which yield, as results, intervals of arbitrarily small width containing, for example, exact values of solutions to differential equations.

At the same time, the computer can be programmed to determine relatively efficient values of the various approximation parameters required by such schemes in order to guarantee desired accuracy.

The use of Taylor series on computers is made practical for a wide class of differential systems by a scheme enabling the computer to derive recursion formulas for the efficient evaluation of Taylor coefficients.

The techniques for automatic error analysis we have presented are for the determination of upper bounds to the overall error including round-off, truncation, initial, accumulated, generated, etc. In fact, our point of view was to consider computations designed to yield intervals known by construction to contain exact solutions. Then the half widths of such intervals are upper bounds to the actual errors in the midpoints, regarding the midpoints as approximate solutions.

In order to study a given source of error in a computational scheme, it is possible to modify some of the procedures we have given by eliminating certain contributions to accumulated interval widths. For example, by modifying the rounded interval arithmetic programs so as to bypass the rounding procedures, we can cause the interval arithmetic computations to ignore round-off errors generated during the course of a computation. On the other hand, we can drop the addition of remainder terms in the truncation of series, etc.

We have tried a number of such modifications of our error bounding procedures. The errors predicted by computations of this sort were, of course, smaller than the strict upper bounds obtained by the rigorous bounding procedures. In many cases they were still actually upper bounds, but in some cases the actual error was much larger than the non-rigorous predicted error. In fact, with non-rigorous error prediction, one can obtain numerical inverses of singular matrices and can integrate past singularities of solutions of differential equations.

13. Acknowledgements

The work described here was supported by the independent research program of Lockheed Missiles and Space Company, Palo Alto,

California. Much of it has appeared in the author's dissertation at Stanford University, 1962.

The major part of the program DIFEQ was written by Ann Davison for the IBM 7094 computer. Without her skill and gracious forbearance throughout the endless revisions and tests requested by the author, this paper would not have been written.

A sincere expression of appreciation goes to George Forsythe for his encouragement of the work from its earliest stages.

As was mentioned in section 9 above, the program for the machine generation of Taylor coefficients, clearly a key part of DIFEQ, is due to the ingenuity of Riley Jaschke.

S. Shayer wrote the interval arithmetic routines used in DIFEQ and kept our program compatible with the computer systems at Lockheed.

The author is grateful for the valuable advice and assistance on various parts of the program, DIFEQ, of R. E. Boche and A. Steigler and for the support and encouragement of many colleagues at Lockheed, especially R. J. Dickson and C. E. Duncan.

REFERENCES

1. Boche, R. E., "An Operational Interval Arithmetic", Presented at IEEE National Electronics Conference, Chicago, Ill., Oct. 1963.

2. Burkill, J. C., "The Derivatives of Functions of Intervals", Fund. Math. 5(1924), pp. 321-327.

3. Burkill, J. C., "Functions of Intervals", Proc. London Math. Soc., 22(1924), pp. 275-336.

4. Collins, G., "Interval Arithmetic for Automatic Error Analysis", M&A-5, Mathematics and Applications Department, IBM, (1960).

5. Davis, H. T., Introduction to Nonlinear Differential and Integral Equations , New York, Dover (1962).

6. Dwyer, P. S., Linear Computations, New York, Wiley (1951).

7. Dwyer, P. S., "Errors of Matrix Computations", Simultaneous Equations and Eigenvalues, Nat. Bur. of Standards, Applied Math. Series 29(1953).

8. Dwyer, P. S., "Matrix Inversion with the Square Root Method", Technometrics, vol. 6, no. 2, (1964), pp. 197-213.

9. Fehlberg, E., "Runge-Kutta Type Formulas of High-order Accuracy and their Application to the Numerical Integration of the Restricted Problem of Three Bodies", International Symposium on Analogue and Digital Techniques Applied to Aeronautics, Liege, Belgium, Sept. 9-12, 1963.

10. Fischer, P. C., "Automatic Propagated and Round-off Error Analysis", 13th National Meeting of the A. C. M., 1958.

11. Fletcher, R., and Reeves, C. M., "A Mechanization of Algebraic Differentiation and the Automatic Generation of Formulae for Molecular Integrals of Gaussian Orbitals", The Computer Journal, v. 6, no. 3, 1963, pp. 287-292.

12. Gibb, A., "Procedures for Range Arithmetic", Algorithm 61, Comm., ACM, 4, 7, July, 1961, 319-320.

13. Gorn, S. and Moore, R. E., "Automatic Error Control- The Initial Value Problem in Ordinary Differential Equations", Aberdeen Proving Ground, Maryland, BRL Report no. 893, March 1953.

14. Gorn, S., "The Automatic Analysis and Control of Computing Errors", J. Soc. Indust. Appl. Math. 2, (1954).

15. Hanson, J. W., Caviness, J. S., and Joseph, C., "Analytic Differentiation by Computer", Comm. ACM, v. 5, no. 1, 1962.

16. Henrici, P., "Discrete Variable Methods in Ordinary Differential Equations", Wiley, New York (1962).

17. Householder, A. S., "Principles of Numerical Analysis", New York, McGraw-Hill (1953).

18. Lowans, A. N., Davids, N., and Levenson, A., "Table of the Zeros of the Legendre Polynomials of Order 1-16 and the Weight Coefficients for Gauss' Mechanical Quadrature Formula", Math. Tables and Other Aids to Comp. (1943).

19. Milne-Thompson, L. M., "The Calculus of Finite Differences", Mac Millan, London, (1933).

20. Moore, R. E., "Automatic Error Analysis in Digital Computation", LMSD-48421, Lockheed Missiles and Space Co., Palo Alto, California, (1959).

21. Moore, R. E., and Yang, C. T., "Interval Analysis", LMSD-285875, Lockheed Missiles and Space Co., Palo Alto, Calif. (1959).

22. Moore, R. E., Strother, W., and Yang, C. T., "Interval Integrals", LMSD-703073, Lockheed Missiles and Space Co., Palo Alto, Calif., (1960).

23. Moore, R. E., "Interval Arithmetic and Automatic Error Analysis in Digital Computing", Applied Math. & Stat. Lab., Stanford University, Technical Report No. 25, (1962).

24. Moore, R. E., Davison, J. A., Jaschke, H. R., and Shayer, S., "DIFEQ Integration Routine - User's Manual", Technical Report LMSC 6-90-64-6, Lockheed Missiles and Space Co., Palo Alto, Calif., (1964).

25. von Neumann, J. and Goldstine, H., "Numerical Inversion of Matrices of High Order", Bull. Amer. Math. Soc., (1947).

26. Rabe, E., "Determination and Survey of Periodic Trojan Orbits in the Restricted Problem of Three Bodies", Astron. Jour., v. 66, (1961), pp. 500-513.

27. Riesz, F. and Sz.-Nagy, B., "Functional Analysis", Ungar, New York (1955), pp. 19ff.

28. Saks, S., "Theory of the Integral", Warsaw (1937), pp. 165-169.

29. Steffensen, J. F., "On the Restricted Problem of Three Bodies", Kgl, Danske Videnskab., Mat.-fys. Medd., vol. 30, (1956), no. 18.

30. Strother, W., "Continuity for Multi-Valued Functions and some Applications to Topology" (doctoral dissertation), Tulane University (1952).

31. Strother, W., "Fixed Points, Fixed Sets, and M-Retracts", Duke Math. J., vol. 22, no. 4(1955), pp. 551-556.

32. Sunaga, T., "Theory of an Interval Algebra and Its Application to Numerical Analysis", RAAG Memoirs II, Gaukutsu Bunken Fukeyu-kai, Tokyo (1958).

33. Young, R. C., "The Algebra of Many-Valued Quantities", Math. Annalen 104 (1931), pp. 260-290.