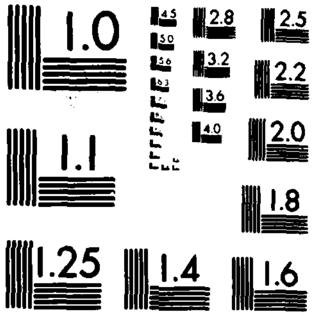
AD-A089 636 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
REPRESENTATIONS OF INTERVALS AND OPTIMAL ERROR BOUNDS.(U)
JUL 80 L B RALL DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2098 NL

END
DATE
FILMED
10 80
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

9 MRC Technical Summary Report #2098

REPRESENTATIONS OF INTERVALS AND OPTIMAL ERROR BOUNDS

L. B. Rall

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

11 July 1980

(Received June 30, 1980)

Sponsored by
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

REPRESENTATIONS OF INTERVALS AND OPTIMAL ERROR BOUNDS

L. B. Rall

Technical Summary Report #2098
July 1980

ABSTRACT

Calculations with interval arithmetic and interval extensions of real functions are often used to obtain lower and upper bounds for what would be the theoretical results of precise computations on exact data. A more traditional way to represent approximate data and results contained in intervals is by means of a representative point x in the interval and an associated measure of error E. For example, the midpoint of an interval has absolute error as an approximation to other points of the interval which is bounded by half the width of the interval. In general, the chosen point is said to be optimal with respect to the measure of error if the maximum value of E over the interval is minimized. A general method for determination of optimal points x^* and minimal error bounds E^* is given. In particular, the harmonic mean of the endpoints of positive intervals is shown to be optimal with respect to relative (or percentage) error, and the geometric mean plays the same role with respect to relative precision, which is a measure of error introduced recently by F. W. J. Olver [SIAM J. Numer. Anal. 15 (1978), 368-393]. In addition, the optimal point and error bound x^*, E^* may be regarded as alternative coordinates for representation of an interval. Explicit rules for interval arithmetic are given in terms of the arithmetic, geometric, and harmonic means of the endpoints and the associated optimal error bounds. Comparisons are also made between the results of exact or rounded interval arithmetic and the *a priori* estimates of relative precision presented in the cited paper by Olver, which show that the intervals resulting from calculation with interval arithmetic can be expected to be smaller than predicted.

AMS (MOS) Subject Classifications: 65G10, 41A50

Key Words: Optimal error bounds, Interval Arithmetic, Absolute error, Relative error, Percentage error, Arithmetic, geometric and harmonic means, Excess width

Work Unit Number 3 (Numerical Analysis and Computer Science)

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041 and the Danish National Science Research Council Grant No. 511-15849.

SIGNIFICANCE AND EXPLANATION

The numerical solution of practical problems almost always involves performing inaccurate calculations on inexactly known data. Thus, in addition to the answers produced by a computer, some indication of their reliability is required. Such an estimate of accuracy may be merely to justify the expense of the computation; however, in cases where the results have implications for human life and safety, as in the design of critical components of aircraft, a guarantee of reliability is imperative. One way to obtain numerical results with such assurance is by interval computation. If the data x are known to lie in an interval I, and one computes an output interval $J = F(I)$, where F is an interval transformation which is an interval extension of a real function f, then the output interval J will contain the values $y = f(x)$ of exact transformations of the data. Ordinarily, interval calculations are done in terms of lower and upper bounds; $I = [a,b]$ is transformed into $J = [c,d]$. The rules for interval calculation are customarily expressed for intervals in this standard form, and in many applications, lower and upper bounds for the exact results are satisfactory. Often, however, one prefers to think of a single number x and its associated error bound E, that is, the maximum error with which x approximates other points of the interval. In particular, x^* is optimal with respect to E if E is minimized on I, the corresponding value E^* being the optimal error bound obtainable. For example, $x^* = m[a,b] = (a+b)/2$ is optimal with respect to absolute error, and $E^* = (b-a)/2$. In other problems, one may wish to optimize different measures of error, such as percentage error or the recently introduced concept of relative precision. A general method for determination of x^*,E^* is given in this paper. For positive intervals, it turns out that the harmonic mean $x^* = h[a,b] = 2ab/(a+b)$ is optimal with respect to percentage error, while the geometric mean $x^* = g[a,b] = \sqrt{ab}$ of the endpoints is optimal with respect to relative precision. The numbers x^*,E^* may be used instead of the endpoints a,b as an alternate representation (or format) for the interval I. Rules for interval arithmetic in some of these alternative formats are given. Finally, it is shown that exact or rounded interval arithmetic produces intervals which are usually smaller than predicted by a priori error estimation, and in no case larger. The results presented are intended to increase the usefulness of interval methods of error estimation.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

REPRESENTATIONS OF INTERVALS AND OPTIMAL ERROR BOUNDS

L. B. Rall

1. Intervals and interval analysis. Just as real and complex numbers are the basic units of real and complex analysis, respectively, the closed finite real intervals

$$I = \{x \mid a \le x \le b\}, \quad -\infty < a \le b < +\infty, \tag{1.1}$$

are the basic units in the branch of mathematics known as interval analysis [4], [5], which is the study of transformations of sets of the type (1.1), which will be called simply intervals for brevity, into others. Geometrically, the set IR of all intervals (1.1) may be visualized as the closed halfplane in the a,b-plane which includes and is bounded below an on the right by the line a = b (see Figure 1).

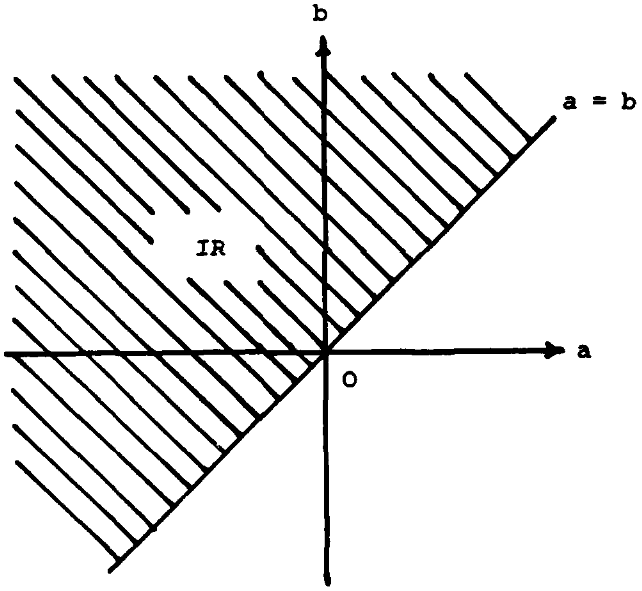


Figure 1. IR as a subset of the a,b-plane.

Thus, IR is two-dimensional. As in the case of complex numbers, which have representations in rectangular coordinates as $z = x + iy$, or in polar coordinates as $z = re^{i\theta}$, various representations are possible for intervals. Some of these representation, which are convenient in one situation or another for computation an error estimation, will be presented in this paper.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041 and the Danish National Science Research Council Grant No. 511-15849.

2. Interval extensions and error esfimation. Interval analysis has important applications to numerical analysis in connection with error estimation. Suppose that one wishes to calculate

(2.1) $$y = f(x),$$

where $f:D\subset R\to R$ is a continuous real function on its domain of definition D. It may happen that x is not known exactly, or is not representable exactly on a given computer; however, it is known that $x\in I$, where I is an interval which can be represented exactly. The set

(2.2) $$f(I) = \{y \mid y = f(x), x\in I\}$$

will be an interval by the continuity of f, but it may not be possible to compute or represent f(I) exactly. For example, f(x) may be defined by an infinite process (series, product, integral, etc.) so that a finite number of computational operations will give an approximation to f(x) which is accurate up to some truncation error, and there will of course be roundoff error in the actual calculations.

Interval analysis deals with the problem of error estimation by the use of interval transformations $F:D\subset IR\to IR$ which are computable (or rounded) interval extensions of real functions f. Thus, one actually computes the interval J defined by

(2.3) $$J = F(I),$$

where F has the following properties:

(i) Representation. F(I) is exactly representable for each representable interval I;

(ii) Extension. F is an extension of f in the sense that $f(I) \subset F(I)$ for each interval I contained in the domain of definition of f;

(iii) Inclusion monotonicity. F is inclusion monotone, which means that for intervals I_1, I_2 in the domain of definition of F, $I_1 \subset I_2 \Rightarrow F(I_1) \subset F(I_2)$.

The interval J may be taken to be the answer to the problem of evaluation of the real transformation (2.1) by inaccurate calculations on inexact data, as it contains the results of all exact values of f for all exact data points $x\in I$.

In traditional numerical analysis, it is customary to think in terms of the error (or more positively, accuracy or precision) of an approximate result, rather than use intervals directly. Thus, one picks a representative point x from the interval, and assigns a measure of error to that point as a way of representing the interval. The interval I is thought of as arising from errors in the data, and the enlargement of f(I) to F(I) as being due to transformation errors (i.e., truncation and roundoff errors, called abbreviation errors by Olver [6]). If the representative point is picked in such a way that the maximum value of the error over the entire interval is minimized, then this point and the resulting error bound are said to be optimal. The representation of intervals by optimal points and error bounds will be illustrated by an example in the next section, following which the general theory will be discussed.

3. An example of an optimal point and error bound. A simple interval method for numerical integration [1] applied to

$$y = \int_0^2 \sqrt{1 + 4x}\, dx \tag{3.1}$$

yields lower and upper bounds for the value of the integral,

$$4.3306334 \leq \int_0^2 \sqrt{1 + 4x}\, dx \leq 4.3452908. \tag{3.2}$$

The standard format for the representation of an interval (1.1) is, of course,

$$I = [a , b] \tag{3.3}$$

in terms of its lower endpoint

$$a = \ell(I) = \min\{x \mid x \epsilon I\}, \tag{3.4}$$

and upper endpoint

$$b = u(I) = \max\{x \mid x \epsilon I\}. \tag{3.5}$$

Thus, (3.2) may be expressed as

$$\int_0^2 \sqrt{1 + 4x}\, dx \ \epsilon\ [4.3306334 , 4.3452908]. \tag{3.6}$$

Another way to write (3.2) is as

$$\int_0^2 \sqrt{1 + 4x}\, dx \simeq 4.3379621 \pm 0.0073287 \tag{3.7}$$

in terms of the midpoint

$$m(I) = m[a , b] = \frac{a + b}{2} \tag{3.8}$$

and the halfwidth (or radius)

$$\alpha(I) = \alpha[a , b] = \frac{b - a}{2} \tag{3.9}$$

of the interval I. This gives

$$I = m(I) \pm \alpha(I) = [m(I) \pm \alpha(I)] \tag{3.10}$$

as an alternative format for the representation of the interval I. In traditional terminology, (3.7) states that one may take m(I) = 4.3379621 as an approximate value for the integral (3.1), with absolute error bounded by $\alpha(I) = 0.0073287$. There are also other measures of error of interest in various problems, such as percentage or relative error, for which the harmonic point of the interval is optimal, or the newly introduced concept of relative precision [6], for which the geometric point plays the same rôle [7]. These lead to alternative representations for intervals. The general situation will now be discussed, after which some examples of useful interval formats and the corresponding rules of interval arithmetic will be given.

4. Optimal points and error estimates. Precise definitions will now be made of the concepts of measures of error and corresponding optimal points and error bounds in intervals.

Definition 4.1. A measure of error of the approximation of a real number y by a real number x is a non-negative continuous function $E = E(x,y)$ which is strictly monotone increasing with respect to the distance

$$d(x,y) = |x - y| \tag{4.1}$$

between x and y and such that

$$E(x,x) = 0. \tag{4.2}$$

Definition 4.2. The error of approximation of points of an interval I by a real number x corresponding to a given measure of error E is

$$\varepsilon(x) = \max_{y \in I}\{E(x,y)\}. \tag{4.3}$$

Definition 4.3. A point $x^* \in I$ is said to be optimal in I with respect to the measure of error E if

$$\varepsilon(x^*) = \min_{x \in I}\{\varepsilon(x)\} = \min_{x \in I}\max_{y \in I}\{E(x,y)\}. \tag{4.4}$$

Although the following result is self-evident, it will be dignified with a proof.

Remark 4.1. The optimal point $x^* \in I$ satisfying (4.4) exists, is unique, and is the solution of the equation

$$E(x,a) = E(x,b). \tag{4.5}$$

Proof: If $a = b$, then $x^* = a = b$; otherwise, the function

$$h(x) = E(x,a) - E(x,b) \tag{4.6}$$

is continuous, and $h(a) < 0$, $h(b) > 0$ by hypothesis. Thus, an $x^* \in I$ exists such that $h(x^*) = 0$, and thus x^* satisfies equation (4.5). Now, suppose that $y < x^*$. It follows that

$$E(y,b) > E(x^*,b) = E(x^*,a) \tag{4.7}$$

as E is strictly monotone increasing in $|x - b|$, and if $y > x^*$, then similarly

$$E(y,a) > E(x^*,a) = E(x^*,b) \tag{4.8}$$

from which the uniqueness and optimality of x^* follow. QED.

Definition 4.4. The value

$$E^* = E(x^*,b) = E(x^*,a) \tag{4.9}$$

obtained for optimal $x^* \epsilon I$ is called the *optimal error estimate* on I for the measure of error E.

One object of this paper is to introduce x^*,E^* as alternative coordinates for the representation of the interval I. As the inverse functions

$$\ell(x,E) = \{y \mid E(x,y) = E,\ y \leq x\}, \tag{4.10}$$

and

$$u(x,E) = \{y \mid E(x,y) = E,\ y \geq x\}, \tag{4.11}$$

are single-valued and continuous by the hypotheses on the error function E, one has the equivalence of the standard representation of the interval I and the x^*,E^* coordinates through the equations

$$a = \ell(x^*,E^*), \quad b = u(x^*,E^*), \tag{4.12}$$

and the definitions of x^*,E^* in terms of a,b by equations (4.5) and (4.9).

5. *The standard format and interval arithmetic*. Interval extensions of the arithmetic operations form the basis of what is called *interval arithmetic* [4], [5]. In the standard format (3.3) for representation of the intervals $I = [a,b]$ and $J = [c,d]$, the fundamental rules for the operations of interval arithmetic are:

(i) *Addition*

$$I + J = [a,b] + [c,d] = [a + b\ ,\ c + d]; \tag{5.1}$$

(ii) *Subtraction*

$$I - J = [a,b] - [c,d] = [a - d\ ,\ b - c]; \tag{5.2}$$

(iii) *Multiplication*

$$I \cdot J = [a,b] \cdot [c,d] = [\min\{ac,ad,bc,bd\}\ ,\ \max\{ac,ad,bc,bd\}]; \tag{5.3}$$

(iv) *Reciprocation* is defined only for intervals J such that $0 \notin J$,

$$J^{-1} = [c\ ,\ d]^{-1} = \left[\frac{1}{d}\ ,\ \frac{1}{c}\right] \text{ if } cd > 0. \tag{5.4}$$

Remark 5.1. Arithmetic operations between real numbers k and intervals are defined by identification of real numbers with *degenerate* intervals; one writes

$$k = [k\ ,\ k] \tag{5.5}$$

and uses the rules for interval arithmetic; for example, if $k \geq 0$, then one has $k\cdot[a\ ,\ b] = [ka\ ,\ kb]$.

Remark 5.2. *Division* is defined as a compound operation; if $0 \notin J$, then

$$I/J = I\cdot J^{-1} \text{ if } c\cdot d = \ell(J)\cdot u(J) > 0. \tag{5.6}$$

Definition 5.1. An interval I is *positive* if $\ell(I) > 0$, negative if $u(I) < 0$, and *zero* if equal to $0 = [0\ ,0]$; I is *positive* (*negative*) *semidefinite* if nonzero, $\ell(I)\cdot u(I) = 0$ and $u(I) > 0$ ($\ell(I) < 0$); finally, I is *indefinite* if $\ell(I)\cdot u(I) < 0$.

Remark 5.3. By analysis of the signs of the intervals I,J, it is possible to calculate the product I·J in certain cases by shortcut methods. However, even with a programmable pocket calculator, it is easy to compute all four products and sort out the smallest and largest. It is also possible to derive explicit (but clumsy) formulas for the product of two intervals, based on the identities

$$\min\{x,y\} = \frac{x + y - |x - y|}{2}, \quad \max\{x,y\} = \frac{x + y + |x - y|}{2}. \tag{5.7}$$

The derivation of such formulas for the product of intervals is left as an exercise for the reader.

Before leaving the familiar standard format for an interval I, it will be noted that this representation provides upper and lower bounds for the points of I, which is the information required in many practical situations. A biologist, for example, may be interested in the range of water temperatures in which a certain species of fish can survive. Another illustration is the "weight and balance" calculation performed by a pilot before flying to determine if the moment of his airplane about its center of gravity along its longitudinal axis is within safe limits for its gross weight. Many other examples of the direct use of intervals may be found, and of problems to which, "The interval is the answer."

6. *The midpoint-halfwidth format and absolute error*. The simplest measure of the error of approximation of y by x is the *absolute error*

$$E(x,y) = |x - y|, \tag{6.1}$$

which is simply the distance between x and y. For this error function, equation (4.5) becomes

$$x - a = b - x, \tag{6.2}$$

which is easily solved for the optimal point and error estimate

$$x^* = (a + b)/2 = m[a\ ,\ b], \tag{6.3}$$

(6.4) $$E^* = (b - a)/2 = \alpha[a\ , b] = \frac{1}{2}\,w[a\ , b],$$

respectively, where the width $w(I) = w[a\ , b]$ of the interval I is of course

(6.5) $$w(I) = w[a\ , b] = b - a.$$

The representation of the interval I in midpoint-halfwidth format (or A-format) is then

(6.6) $$I = [m(I) \pm \alpha(I)],$$

where the comma in the standard format has been replaced by the ± sign. (One can also use the notation (6.6) without the square brackets, if no confusion is entailed.) As has been shown above (and has been known for a long time), this representation is optimal with respect to absolute error. The transformation from (6.6) back to the standard format is by means of the evident relationships

(6.7) $$a = \ell(I) = m(I) - \alpha(I), \quad b = u(I) = m(I) + \alpha(I).$$

The basic rules for interval arithmetic in this format are:

(i) Addition

(6.8) $$m(I + J) = m(I) + m(J), \quad \alpha(I + J) = \alpha(I) + \alpha(J);$$

(ii) Subtraction

(6.9) $$m(I - J) = m(I) - m(J), \quad \alpha(I - J) = \alpha(I) + \alpha(J);$$

(iii) Multiplication

(6.10)
$$\begin{aligned} m(I\cdot J) &= m(I)\cdot m(J) + \tfrac{1}{2}\{|\alpha(I)\cdot m(J) + A(I)\cdot\alpha(J)| - \\ &\qquad - |\alpha(I)\cdot m(J) - A(I)\cdot\alpha(J)|\}, \\ \alpha(I\cdot J) &= \mu(I)\cdot\alpha(J) + \tfrac{1}{2}\{|\alpha(I)\cdot m(J) + A(I)\cdot\alpha(J)| + \\ &\qquad + |\alpha(I)\cdot m(J) - A(I)\cdot\alpha(J)|\}, \end{aligned}$$

where

(6.11)
$$\begin{aligned} A(I) &= \tfrac{1}{2}(|m(I) + \alpha(I)| - |m(I) - \alpha(I)|), \\ \mu(I) &= \tfrac{1}{2}(|m(I) + \alpha(I)| + |m(I) - \alpha(I)|). \end{aligned}$$

(iv) Reciprocation is defined only for intervals J such that $0 \notin J$,

(6.12)
$$\begin{aligned} m(J^{-1}) &= m(J)/(m(J)^2 - \alpha(J)^2), \\ \alpha(J^{-1}) &= \alpha(J)/(m(J)^2 - \alpha(J)^2), \text{ if } m(J)^2 > \alpha(J)^2. \end{aligned}$$

Remark 6.1. There are several alternative expressions to (6.10)-(6.11) for the product of two intervals in midpoint-halfwidth format, which the reader may wish to derive.

Remark 6.2. The condition $m(J)^2 > \alpha(J)^2$ in (6.12) enforces $|m(J)| > \alpha(J)$, and thus the interval J will be positive or negative.

Remark 6.3. The midpoint m[a , b] is also called the average, or arithmetic mean of the numbers a,b.

Finally, the set IR of intervals may be represented in the m,α coordinate system simply as the closed upper halfplane of the m,α-plane (see Figure 2).

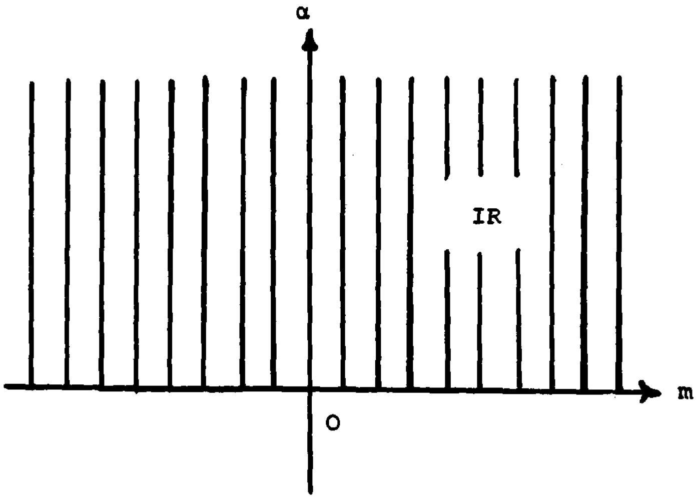


Figure 2. IR as a subset of the m,α-plane.

7. The harmonic point-relative width format. Relative and percentage error.

A measure of error which is useful in many circumstances is the relative error of approximation of a real number y ≠ 0 by a real number x of the same sign, which is defined by

$$(7.1) \qquad E(x,y) = \left|\frac{x - y}{y}\right|.$$

For simplicity, it will be assumed that the numbers being considered are positive; the same results hold, with appropriate changes of sign, for intervals of negative numbers. For relative error, equation (4.5) becomes

$$(7.2) \qquad \frac{x - a}{a} = \frac{b - x}{b},$$

which has as its solution

$$(7.3) \qquad x^* = \frac{2ab}{a + b} = h[a , b] = h(I),$$

the harmonic point of I (also called the harmonic mean of a,b), with the corresponding optimal error estimate

$$(7.4) \qquad E^* = \frac{b - a}{a + b} = r[a , b] = r(I),$$

the *relative width* of I. This gives the corresponding *harmonic point-relative width* format (or *R-format*) for the representation of the interval I:

(7.5) $$I = [h(I)\ R\ r(I)].$$

The transformation from this format back to the standard format is given by

(7.6) $$a = \ell(I) = h(I)/(1 + r(I)), \quad b = u(I) = h(I)/(1 - r(I)).$$

For positive intervals I,J, the basic rules of interval arithmetic in the R-format are:

(i) *Addition*

(7.7)
$$h(I + J) = \frac{(h(I) + h(J))^2 - (h(I)r(J) + h(J)r(I))^2}{h(I) + h(J) - (h(I)r(J)^2 + h(J)r(I)^2)},$$

$$r(I + J) = \frac{h(I)r(I) + h(J)r(J) - (h(I)r(I)r(J)^2 + h(J)r(J)r(I)^2)}{h(I) + h(J) - (h(I)r(J)^2 + h(J)r(I)^2)};$$

(ii) *Subtraction* is defined only if I - J is positive,

(7.8)
$$h(I - J) = \frac{(h(I) - h(J))^2 - (h(I)r(J) + h(J)r(I))^2}{h(I) - h(J) - (h(I)r(J)^2 + h(J)r(I)^2)},$$

$$r(I - J) = \frac{h(I)r(I) + h(J)r(J) - (h(I)r(I)r(J)^2 + h(J)r(J)r(I)^2}{h(I) - h(J) - (h(I)r(J)^2 + h(J)r(I)^2)},$$

if $h(I) - h(J) > h(I)r(J) + h(J)r(I)$;

(iii) *Multiplication*

(7.9) $$h(I \cdot J) = \frac{h(I)h(J)}{1 + r(I)r(J)}, \quad r(I \cdot J) = \frac{r(I) + r(J)}{1 + r(I)r(J)};$$

(iv) *Reciprocation*

(7.10) $$h(J^{-1}) = \frac{1 - r(J)^2}{h(J)}, \quad r(J^{-1}) = r(J).$$

Remark 7.1. The condition on subtraction in (7.8) insures that $d < a$ in the standard format $I = [a\ ,b]$, $J = [c\ ,d]$.

Remark 7.2. The formulas (7.7)-(7.8) for addition and subtraction in this format are unwieldy; a transformation to and from the standard format is probably preferable for these operations.

Example 7.1. The interval expression (3.6) may be written

(7.11) $$\int_0^2 \sqrt{1 + 4x}\, dx \in [4.3379497\ R\ 0.001689438]$$

in the R-format. This means that one may take $x^* = h(I) = 4.3379497$ as the approximate value of the integral (3.1), with relative error bounded by $E^* = r(I) = 0.001689438$, which can be symbolized by

(7.12) $$\int_0^2 \sqrt{1 + 4x}\, dx \simeq 4.3379497 \text{ R } 0.001689438.$$

Remark 7.3. In conversion from $[a\,, b]$ to $[h(I) \text{ R } r(I)]$ format, $r(I)$ is rounded upward, if necessary, so that $[a\,, b] \subset [h(I) \text{ R } r(I)]$. This *directed rounding* has been employed in (7.11) and the numerical examples cited below.

Remark 7.4. The *percentage error* is 100 times the relative error (7.1), so that $x^* = h(I)$ is also optimal for percentage error, with

(7.13) $$E^* = p(I) = 100 \cdot r(I)$$

being the corresponding optimal error bound.

Thus, one may write (7.12) as

(7.14) $$\int_0^2 \sqrt{1 + 4x}\, dx \simeq 4.3379497 \pm 0.1689438\%,$$

showing that the percentage error in taking $h(I) = 4.3379497$ as an approximation to y in this example is less than 0.17%. The *percentage format* (or *P-format*) for representation of positive intervals is thus

(7.15) $$I = h(I) \pm p(I)\% = [h(I) \,\%\, p(I)].$$

8. *The geometric point-ratio format. Relative precision and approximate relative precision.* The *geometric point* $g(I) = g[a\,, b] = \sqrt{ab}$ and the *ratio* $\rho(I) = \rho[a\,, b] = \sqrt{b/a}$ may also be used to represent a positive interval $I = [a\,, b]$. This coordinate system will be called the *geometric point-ratio format* (or *G-format*), and symbolized by

(8.1) $$I = [g(I) \divideontimes \rho(I)].$$

The sign $\divideontimes$ is a combination of the × and ÷ signs, symbolizing the transformation

(8.2) $$a = \ell(I) = g(I)/\rho(I), \quad b = u(I) = g(I) \cdot \rho(I)$$

from the format (8.1) into the standard format, in the same way that ± is used in the midpoint-halfwidth format. The geometric point $g(I) = g[a,b] = \sqrt{ab}$ is also called the *geometric mean* of the positive numbers a,b.

The G-format is related to error estimation through the *maximum ratio function* M defined by

(8.3) $$M(x,y) = \max\left\{\frac{x}{y}, \frac{y}{x}\right\} = \frac{x^2 + y^2 + |x^2 - y^2|}{2xy}$$

for $x > 0$, $y > 0$. On the positive interval $I = [a\,, b]$, the minimum of the maximum value of M is attained for x^* satisfying

(8.4) $$\frac{x}{a} = \frac{b}{x},$$

that is, $x^* = g(I)$. This point will be optimal for each error function

(8.5) $$E(x,y) = f(M(x,y)),$$

where f is continuous, strictly monotone increasing, and $f(1) = 0$. As

(8.6) $$M(x^*,a) = M(x^*,b) = \rho(I),$$

one has

(8.7) $$E^* = f(\rho(I))$$

as the optimal error estimate for the error function (8.5). In particular, the <u>relative</u> <u>precision</u>

(8.8) $$\text{r.p.}(I) = \ln \rho(I)$$

and <u>approximate</u> <u>relative</u> <u>precision</u>

(8.9) $$\text{a.r.p.}(I) = \rho(I) - 1$$

defined by Olver [6] fit into this category, and lead to the respective formats

(8.10) $$I = [g(I)\ \text{r.p.}\ \ln \rho(I)] \text{ and } I = [g(I)\ \text{a.r.p.}\ \rho(I) - 1]$$

for the representation of intervals.

Example 8.1. In the formats (8.1) and (8.10), the interval expression (3.6) may be written

(8.11) $$\int_0^2 \sqrt{1 + 4x}\, dx \in [4.3379559 \ast 1.0016909],$$

or as

(8.12) $$\int_0^2 \sqrt{1 + 4x}\, dx \simeq 4.3379559\ \text{r.p.}\ 0.001689438,$$

or

(8.13) $$\int_0^2 \sqrt{1 + 4x}\, dx \simeq 4.3379559\ \text{a.r.p.}\ 0.0016909.$$

Directed rounding has been used in the conversion of (3.6) into the above formats.

As conversion between the r.p. and a.r.p. formats (8.10) and the G-format (8.1) is immediate, the rules of interval arithmetic will be stated for the G-format.

(i) <u>Addition</u>

(8.14) $$g(I + J)^2 = (g(I) + g(J))^2 - \frac{g(I)g(J)}{\rho(I)\rho(J)}(\rho(I) - \rho(J))^2,$$
$$\rho(I + J)^2 = \rho(I)\rho(J)\frac{g(I)\rho(I) + g(J)\rho(J)}{g(I)\rho(J) + g(J)\rho(I)};$$

(ii) Subtraction is defined only if I - J is positive,

$$g(I - J)^2 = (g(I) - g(J))^2 - \frac{g(I)g(J)}{\rho(I)\rho(J)}(\rho(I)\rho(J) - 1)^2,$$

(8.15)

$$\rho(I - J)^2 = \frac{\rho(I)}{\rho(J)}\left(\frac{g(I)\rho(I)\rho(J) - g(J)}{g(I) - g(J)\rho(I)\rho(J)}\right) \text{ if } g(I) - g(J)\rho(I)\rho(J) > 0;$$

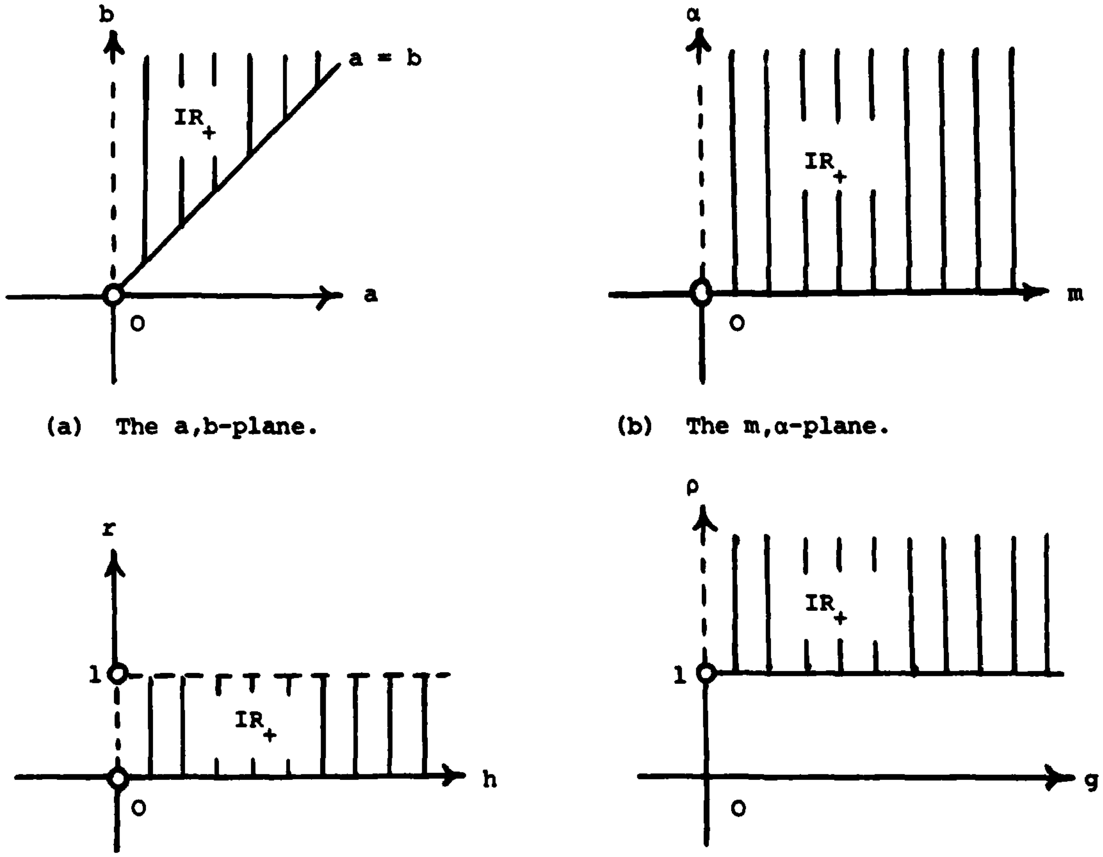
(iii) Multiplication

(8.16) $g(I \cdot J) = g(I) \cdot g(J), \quad \rho(I \cdot J) = \rho(I) \cdot \rho(J);$

(iv) Reciprocation

(8.17) $g(J^{-1}) = 1/g(J) = g(J)^{-1}, \quad \rho(J^{-1}) = \rho(J).$

The set IR_+ of positive intervals is shown in Figure 3 in the [a , b], [m ± α], [h R r], and [g * ρ] coordinate systems.



(a) The a,b-plane.

(b) The m,α-plane.

(c) The h,r-plane.

(d) The g,ρ-plane.

Figure 3. The set IR_+ of positive intervals.

Remark 8.1. The transformations from the G-format into the relative precision (r.p.) and approximate relative precision (a.r.p.) formats may be made by using the formulas

$$(8.18) \qquad \rho(I) = e^{r.p.(I)},$$

and

$$(8.19) \qquad \rho(I) = 1 + a.r.p.(I),$$

which follow directly from (8.8) and (8.9), respectively. From (8.14)-(8.17), the rules of interval arithmetic in the r.p. format are obtained from the following formulas for the relative precision, the geometric point being given in the same way.

(i) Addition

$$(8.20) \qquad r.p.(I + J) = \frac{r.p.(I) + r.p.(J)}{2} + \frac{1}{2}\ln\{\frac{g(I)\rho(I) + g(J)\rho(J)}{g(I)\rho(J) + g(J)\rho(I)}\};$$

(ii) Subtraction is defined only if I - J is positive,

$$(8.21) \qquad \begin{array}{l} r.p.(I - J) = \dfrac{r.p.(I) - r.p.(J)}{2} + \dfrac{1}{2}\ln\{\dfrac{g(I)\rho(I)\rho(J) - g(J)}{g(I) - g(J)\rho(I)\rho(J)}\}, \\ \text{if } g(I) - g(J)\rho(I)\rho(J) > 0; \end{array}$$

(iii) Multiplication

$$(8.22) \qquad r.p.(I \cdot J) = r.p.(I) + r.p.(J);$$

(iv) Reciprocation

$$(8.23) \qquad r.p.(J^{-1}) = r.p.(J).$$

Remark 8.2. The $[g * \rho]$ format is related to the $[m \pm \alpha]$ format for positive intervals by means of logarithms. If I is positive, then the logarithm of I is defined to be the interval

$$(8.24) \qquad \ln(I) = [\ln\{\ell(I)\}, \ln\{u(I)\}].$$

Thus,

$$(8.25) \qquad \ln\{g(I)\} = \frac{1}{2}(\ln\{\ell(I)\} + \ln\{u(I)\}) = m(\ln(I)),$$

and

$$(8.26) \qquad \ln\{\rho(I)\} = r.p.(I) = \frac{1}{2}(\ln\{u(I)\} - \ln\{\ell(I)\}) = \alpha(\ln(I)),$$

as noted by Olver [6].

9. Choice of format for intervals. In various applications, one of the formats presented above may be preferable to the others. For error estimation in positive intervals, the classical inequality between the arithmetic, geometric, and harmonic means of two positive numbers a,b is

$$(9.1) \qquad h[a, b] \le g[a, b] \le m[a, b],$$

with equality if and only if a = b [2]. Thus, the choice of $x^* = g(I)$ to optimize relative precision may be viewed as a compromise between minimization of absolute and relative errors over I. In addition, the choice of the geometric point has important theoretical and computational implications, as discussed by Olver [6].

For expression of interval arithmetic, the standard and the midpoint-halfwidth representations are simpler for addition and subtraction, while the harmonic point-relative width and geometric point-ratio formats are simpler for multiplication and reciprocation (and hence division). It should be pointed out that for positive intervals,

(9.2) $$I \cdot J = [ac, bd]$$

in standard format, and for the A-format, it follows from (6.11) that $A(I) = \alpha(I)$, $\mu(I) = m(I)$, so that (6.10) simplifies to

(9.3) $$\begin{aligned} m(I \cdot J) &= m(I) \cdot m(J) + \alpha(I) \cdot \alpha(J), \\ \alpha(I \cdot J) &= m(I) \cdot \alpha(J) + \alpha(I) \cdot m(J). \end{aligned}$$

These formulas compare favorably in simplicity to the corresponding rules in the R- and G-formats.

Rather than insisting on one format or another throughout a computation, it may be expedient to transform from one to another during the calculation, or to represent input data or output results. Relations such as

(9.4) $$m(J^{-1}) = 1/h(J) = h(J)^{-1}, \quad \alpha(J^{-1}) = \alpha(J)/g(J)^2,$$

which are among the many charming relationships between the arithmetic, geometric, and harmonic means, indicate the possiblities of coordinate transformations in interval calculations.

10. A priori error estimates and excess width. At any stage in an interval computation, one has an output interval $J = F(I)$ for which the representative point $x^*(J)$ and the optimal error estimate (or bound) $E^*(J)$ for a given measure of error E may be found by expressing J in x^*, E^* coordinates. This is called a posteriori error estimation, as the calculations called for by the transformation F have been performed. Another type of error analysis is concerned with the estimation of $x^*(J)$, $E^*(J)$ (usually $E^*(J)$ in particular), given $x^*(I)$, $E^*(I)$ for the input interval I, before the transformation F is performed to obtain J. This is called a priori error estimation. From the standpoint of interval analysis, a priori error analysis gives an estimated interval K such that $J \subset K$, and hence

(10.1) $$E^*(K) \geq E^*(J),$$

with the inequality being strict if the excess width of K over its subinterval is positive, where excess width is defined to be

$$d(K,J) = \max\{\ell(J) - \ell(K)\ ,\ u(K) - u(J)\} \quad \text{for } J \subset K, \tag{10.2}$$

which is simply the distance between K and J in the ordinary metric topology for intervals [4], [5].

<u>A priori</u> error estimation may be viewed as a type of approximate interval arithmetic, in which the results of various operations are expressed by simple formulas which yield intervals containing the results which would be actually computed. The highly interesting methods of Olver [6] may be interpreted in this light. For example ([6], (2.2)), the expression

$$a \simeq \tilde{a};\ \mathrm{ap}(\alpha), \tag{10.3}$$

which means that $\tilde{a}$ is an approximation to a of <u>absolute precision</u> α in the terminology of [6], may be expressed in the notation of the present paper as

$$a \in [\tilde{a} \pm \alpha] = I, \tag{10.4}$$

i.e., $\tilde{a} = m(I)$, $\alpha = \alpha(I)$. Olver gives ([6], (2.4)) the formula

$$ab \simeq \widetilde{ab};\ \mathrm{ap}(|\tilde{a}|\beta + |\tilde{b}|\alpha + \alpha\beta), \tag{10.5}$$

which in interval notation defines the interval K with

$$\begin{aligned} m(K) &= m(I)m(J), \\ \alpha(K) &= |m(I)|\alpha(J) + |m(J)|\alpha(I) + \alpha(I)\alpha(J), \end{aligned} \tag{10.7}$$

where $J = [\tilde{b} \pm \beta]$ contains the point b. Comparison of (10.7) with (9.3) shows that (10.5) represents an approximation K to I•J which, for I,J positive, has excess width

$$d(K\ ,\ I\cdot J) = 2\alpha(I)\alpha(J), \tag{10.8}$$

which is positive if I,J are each nondegenerate. It may be noted that u(K) = u(I•J), i.e., formula (10.7) gives the correct <u>upper</u> endpoint of I•J for positive I,J.

Similarly, Olver's definition ([6], (3.2)) of the <u>relative precision</u> of approximation of a by $\tilde{a}$,

$$a \simeq \tilde{a};\ \mathrm{rp}(\alpha), \tag{10.9}$$

means

$$a \in [\tilde{a} \divideontimes e^{\mathrm{rp}(\alpha)}] = I \tag{10.10}$$

in the notation of the present paper, that is, $\tilde{a} = g(I)$, $\mathrm{rp}(\alpha) = \ln \rho(I)$. For relative precision of addition ([6], (3.5)), Olver gives the formula

$$(10.11) \qquad a + b \simeq \tilde{a} + \tilde{b}; \quad rp(\ln\{\frac{\tilde{a}e^{\alpha} + \tilde{b}e^{\beta}}{\tilde{a} + \tilde{b}}\}).$$

In the notation of §8, this defines an interval K with

$$(10.12) \qquad g(K) = g(I) + g(J), \quad \rho(K) = \frac{g(I)\rho(I) + g(J)\rho(J)}{g(I) + g(J)} .$$

In standard format, the interval I + J is

$$(10.13) \qquad I + J = [\frac{g(I)}{\rho(I)} + \frac{g(J)}{\rho(J)}, g(I)\rho(I) + g(J)\rho(J)],$$

which shows that u(K) = u(I + J), and K has excess width

$$(10.14) \qquad d(K, I + J) = \frac{g(I)g(J)(\rho(I) - \rho(J))^2}{\rho(I)\rho(J)(g(I)\rho(I) + g(J)\rho(J))} ,$$

which will be positive if $\rho(I) \neq \rho(J)$.

The error analysis given by Olver in [6] also provides formulas to predict the results obtained with various types of rounding. As adapted to directed rounding used in interval arithmetic, excess width in the predicted intervals will also be observed as above. This does not detract from the usefulness of the a priori methods given in [6]; it simply means that one can expect the results obtained by the use of properly rounded interval arithmetic to be better than predicted.

11. Historical remarks. The optimality of the midpoint of an interval with respect to absolute error has been known for a long time. The same relationship between the harmonic point and the relative (or percentage) error was learned from George Pólya in a seminar he conducted at the University of California, Berkely, in 1956. On the basis of these results, the A-, R-, and P-formats were made available as optional alternatives to the standard format in the programs [1], [3], for interval numerical integration and the solution of nonlinear systems of equations, respectively. The results given above on the significance of the geometric point for relative precision had to wait, of course, for the fundamental paper by Olver [6].

12. Acknowlegments. This paper was written during a leave at the Institute for Applied Mathematics, University of Freiburg, Germany. The gracious host of the Author was Prof. Dr. Karl Nickel, founder of the Interval Library, a unique and effective research tool.

References

1. Julia H. Gray and L. B. Rall. INTE: A UNIVAC 1108/1110 program for numerical integration with rigorous error estimation, Mathematics Research Center, MRC Tech. Rep. 1428, University of Wisconsin-Madison, October, 1975.
2. G. H. Hardy, J. E. Littlewood, and G. Pólya. Inequalities, 2nd Ed., Cambridge University Press, Cambridge, England, 1952.
3. Dennis Kuba and L. B. Rall. A UNIVAC 1108 program for obtaining rigorous error estimates for approximate solutions of systems of equations, Mathematics Research Center, MRC Tech. Rep. 1168, University of Wisconsin-Madison, January, 1972.
4. R. E. Moore. Interval Analysis, Prentice-Hall, Englewood Cliffs, N. J., 1966.
5. R. E. Moore. Methods and Applications of Interval Analysis, SIAM Publications, Philadelphia, Pa., 1979.
6. F. W. J. Olver. A new approach to error arithmetic, SIAM J. Numer. Anal. 15 (1978), 368-393.
7. L. B. Rall. Applications of software for automatic differentiation in numerical computation, Computing, Suppl. 2 (1980), 141-156.

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2098	2. GOVT ACCESSION NO. AD-A089 636	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* REPRESENTATIONS OF INTERVALS AND OPTIMAL ERROR BOUNDS		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) L. B. Rall		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041 511-15849
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 3 - Numerical Analysis and Computer Science
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE July 1980
		13. NUMBER OF PAGES 17
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Optimal error bounds, Interval Arithmetic, Absolute error, Relative error, Percentage error, Arithmetic, geometric and harmonic means, Excess width

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Calculations with interval arithmetic and interval extensions of real functions are often used to obtain lower and upper bounds for what would be the theoretical results of precise computations on exact data. A more traditional way to represent approximate data and results contained in intervals is by means of a representative point x in the interval and an associated measure of error E. For example, the midpoint of an interval has absolute error as an approximation to other points of the interval which is bounded

(continued)

ABSTRACT (cont.)

by half the width of the interval. In general, the chosen point is said to be optimal with respect to the measure of error if the maximum value of E over the interval is minimized. A general method for determination of optimal points x* and minimal error bounds E* is given. In particular, the harmonic mean of the endpoints of positive intervals is shown to be optimal with respect to relative (or percentage) error, and the geometric mean plays the same role with respect to relative precision, which is a measure of error introduced recently by F. W. J. Olver [SIAM J. Numer. Anal. 15 (1978), 368-393]. In addition, the optimal point and error bound x*, E* may be regarded as alternative coordinates for representation of an interval. Explicit rules for interval arithmetic are given in terms of the arithmetic, geometric, and harmonic means of the endpoints and the associated optimal error bounds. Comparisons are also made between the results of exact or rounded interval arithmetic and the *a priori* estimates of relative precision presented in the cited paper by Olver, which show that the intervals resulting from calculation with interval arithmetic can be expected to be smaller than predicted.