Interval Arithmetic: From Principles to Implementation

T. HICKEY AND Q. JU

Brandeis University, Waltham, Massachusetts

AND

M. H. VAN EMDEN

University of Victoria, Victoria, B.C., Canada

Abstract. We start with a mathematical definition of a real interval as a closed, connected set of reals. Interval arithmetic operations (addition, subtraction, multiplication, and division) are likewise defined mathematically and we provide algorithms for computing these operations assuming exact real arithmetic. Next, we define interval arithmetic operations on intervals with IEEE 754 floating point endpoints to be sound and optimal approximations of the real interval operations and we show that the IEEE standard's specification of operations involving the signed infinities, signed zeros, and the exact/inexact flag are such as to make a correct and optimal implementation more efficient. From the resulting theorems, we derive data that are sufficiently detailed to convert directly to a program for efficiently implementing the interval operations. Finally, we extend these results to the case of general intervals, which are defined as connected sets of reals that are not necessarily closed.

Categories and Subject Descriptors: D.2.4 [**Software Engineering**]: Software/Program Verification—*correctness proofs*; *validation*; G.0 [**Mathematics of Computing**]: General; G.1.0 [**Numerical Analysis**]: General—*computer arithmetic*; *conditioning* (*and ill-conditioning*); *error analysis*; *interval arithmetic*; G.1.5 [**Numerical Analysis**]: Roots of Nonlinear Equations—*iterative methods*

General Terms: Algorithms, Reliability, Theory, Verification

1. *Introduction*

One of the earliest lessons of digital numerical computation was that, although most programs give highly accurate results, it can happen that rounding errors build up in such a way that none of the many decimals in the result is meaningful. A good early summary of this is Forsythe's [1970] "Pitfalls of Computation." Numerical analysis emerged as the science of determining conditions under which algorithms give accurate results.

Instead of using a single floating-point number as approximation for the value of a real variable in the mathematical model under investigation, interval arithmetic

Authors' addresses: T. Hickey and Q. Ju, Department of Computer Science, Brandeis University, Waltham, MA 02254, e-mail: tim@cs.brandeis.edu; http://www.cs.brandeis.edu/~tim; M. H. van Emden, Department of Computer Science, University of Victoria, Victoria, B.C., Canada, e-mail:vanemden@csr.uvic.ca; http://www.csr.uvic.ca/~vanemden.

acknowledges limited precision by associating with the variable *a set of reals* as possible values. For ease of storage and computation, these sets are restricted to intervals. The computation rules aim at maintaining the property of containing all possible values.

Initially, the attraction of interval arithmetic was that it would not be necessary to analyze whether the conventional pointwise floating-point computations are safe. As the interval results contain all possible values, a narrow interval indicates success. A wide interval does not prove that a conventionally computed result is wrong, but it does indicate a risk.

When evaluating expressions in interval arithmetic, the interval result tends to be disappointingly large. This is due partly to its conservative nature: the result has to contain all possible values, including those where rounding errors combine in an unfavourable way. But the main cause of wide intervals is that interval arithmetic does not identify different occurrences of the same variable. Because it treats these as if they were unique occurrences, the result is much wider than one could reasonably expect.

Interval arithmetic has the property of *correctness*: result intervals are guaranteed to contain the real number that is the value of the expression. Implementations can only realize this potential by rounding floating-point operations in the right direction. In the early days, such code was not portable, if rounding direction could be controlled at all. In addition, in the early days of interval arithmetic, the extra demands of interval computation on the processor and on memory (several operations instead of one; two numbers to be stored instead of one) posed serious problems.

The utility of interval algorithms has shifted over time from automatically performing rigorous error analysis to solving nonlinear problems, including: systems of nonlinear equations, global optimization (unconstrained and constrained), and integrating differential equations. One of the keys to overcoming overly conservative interval bounds, even in the presence of rounding, is an algorithm that is a contracting map. A contracting map produces a sequence of interval results that are successively narrower subintervals of each other. With infinite precision interval arithmetic, the final result width of a contracting map can approach zero. In practice, with finite-precision arithmetic and positive-width interval parameters, a point of diminishing returns is reached, a point at which finite precision prevents further contraction of the result interval.

Most expositions of interval arithmetic require intervals to be bounded from above and below. They disallow unbounded intervals and they disallow division of intervals when the denominator contains zero. In practice, however, these classical restrictions on interval operations must sometimes be subverted. The restriction of division to interval denominators not containing zero means that the interval division operator, like its real counterpart, is a partial function, and this lack of totality creates the same kinds of problems as arise for pointwise computation. Tests have to be inserted for the presence of zero in the divisor interval, which then needs to be split. On the other hand, a correct and total interval division operation allows the algorithm to proceed independently of whether the divisor interval contains zero or not.

In this paper, we present a system of interval arithmetic that has the following properties:

(1) Correctness
(2) Totality
(3) Closedness
(4) Optimality
(5) Efficiency

Contributions to these goals are scattered over a number of publications. Looking at any one of these, one may get the impression that the state of the art is far from achieving these goals. But, as we show in this paper, relatively little needs to be added to the combined literature. We now comment on each of the above five criteria.

Correctness. The criterion for correctness of a definition of interval arithmetic is that the "Fundamental Theorem of Interval Arithmetic" hold: when an expression is evaluated using intervals, it yields an interval containing all results of pointwise evaluations based on point values that are elements of the argument intervals. Existing proofs of this restrict the types of interval arithmetic definitions for which it can be used. To minimize such a restriction, we give a new and more general proof that is based on elementary facts of set theory.

Totality. A partial function is said to be *total* if it is defined for all possible arguments. This is a desirable property for arithmetic operations, which is lacking in division on the reals. This lack of totality gives rise to many problems in non-interval arithmetic. In the literature, interval division is often not defined when the divisor interval contains zero, thus inheriting and even aggravating this problem in non-interval arithmetic. We follow those treatments in the interval literature where the interval operators ($+,-,*,/$) are defined on all intervals. The resulting sets will not always be intervals, but will in all cases be a finite union of intervals.

Closedness. It is desirable that an operation on intervals yields an interval. In conventional mathematical terminology, this says that the set of intervals be *closed* under the operation. The arithmetic interval operations are closed on the set of finite unions of general intervals, but it may be prohibitively expensive to work on this domain as the number of sets in the union can grow rapidly. Another way to obtain a closed system of interval arithmetic is to compose the interval operators as defined here with an operation that maps a set into the smallest interval containing it. This latter operation can be performed after each individual interval operation, or after evaluating an entire expression. In either case, the closedness property is maintained.

Optimality. By optimality, we mean that the computed floating-point interval is not wider than necessary. In some cases, the difference that optimality makes is so small as to require a particular endpoint of an interval to be open rather than closed. We show that for primitive arithmetic operations ($+,-,*,/$), if exact real arithmetic is used, our algorithm computes the image of the arithmetic operators over connected sets of reals.

Although optimality is lost for expressions with repeated variables, there is a plethora of classical techniques for transforming expressions so as to minimize the overestimation resulting from interval evaluation. Those techniques transfer to our system.

Efficiency. Slow execution has been one of the reasons for the early lack of acceptance of interval arithmetic. Sometimes, interval arithmetic operations are effected by subroutine calls. In this case, elimination of the calling overhead is likely to be the main source possible improvement. Other possibilities for speed-up, which we consider in this paper, arise from exploiting certain details of floating-point arithmetic.

Now, and for some time to come, numerical computation is done by floating-point arithmetic on IEEE 754 standard processors. To avoid loss of speed, interval computation should make use of the possibility that operations involving one of the infinities can be performed at the same speed as the other operations. Loss of speed also arises from the need to prevent undefined values by means of run-time tests. On pipelined architectures, tests can give significant delays by causing the pipeline to be emptied.

Our article includes tables that make it possible to execute interval arithmetic code without including any tests for infinite endpoints; for example, our formulas allow one to avoid subtraction of infinities of the same sign. They also ensure that division by zero yields the infinity of the right sign by a single invocation of the standard floating-point division instruction. Thus, we avoid run-time tests by suitably structuring the code.

The algorithms in this paper have been used to implement several logically sound constraint solvers based on interval arithmetic [Hickey 2000a, 2000b; Hickey et al. 2000].

2. *Semantical Treatment of Real Intervals and Their Operations*

The literature is unanimous in the treatment of bounded intervals as sets of reals: in $[a, b]$, a and b are finite floating-point numbers and the meaning of the expression $[a, b]$ is the set $\{x \in \mathcal{R} \mid a \leq x \leq b\}$. This is the common interpretation of the expression $[a, b]$ as a "closed" interval: one that contains its endpoints. Things become less obvious if one wants to avoid exceptions and allow division by intervals containing zero. In such cases, the result is an unbounded interval, which happens to have a convenient representation by allowing a or b to be one of the infinities provided by the IEEE 754 standard. The question then arises whether the interval is still "closed." There are also disagreements on the meaning of difficult cases such as $[0, 0]/[0, 0]$, or even whether such cases should be defined at all.

As a first step in avoiding these difficulties, we distinguish between *syntax* and *semantics* of intervals. Syntax considers expressions. Semantics determines the mathematical objects denoted by the expressions.

Semantically, we show that the ambitious goals described in the introduction can be achieved by regarding intervals as sets of reals. As these sets can be unbounded, we need to be careful about the meaning of "closed." For other reasons also, it is important to review certain concepts of topology, as these allow us to give the most succinct and precise characterization of the semantics of intervals.

Definition 1. A *basic open set* of reals is a set of the form $\{x \in \mathcal{R} \mid a < x < b\}$ where a, b are real numbers. A set S of reals is *open* if, for every point x in S, there is a basic open set U_x such that $x \in U_x \subset S$. A set S of reals is *closed* if its complement is open. A set S of reals is *connected* if there do not exist disjoint non-empty open sets U_1 and U_2 that each intersect S and for which $S \subset U_1 \cup U_2$.

THEOREM 1. *Let a and b be reals. The following are closed connected sets of reals*:

$$\{x \in \mathcal{R} \mid a \leq x \leq b\}, \{x \in \mathcal{R} \mid x \leq b\},$$
$$\{x \in \mathcal{R} \mid a \leq x\}, \textit{ and } \mathcal{R}.$$

There are no other closed connected sets of reals.

This theorem is a well-known result in topology; see, for example, Lipschutz [1965]. Note that the theorem also identifies ∅ as a closed, connected set.

Definition 2. A *real interval* is a closed connected set of reals.

In the sequel, we introduce a special notation for real intervals. We define this notation only for $a \leq b$ and we use a noninterval notation, ∅, for the real interval that is the empty set.

2.1. ARITHMETIC OPERATIONS ON REAL INTERVALS.

The literature is unanimous in defining the interval operations $X + Y$, $X - Y$, and $X * Y$ by

Definition 3. Let X and Y be real intervals, then their sum, difference, and product are defined to be the following sets:

$$\begin{aligned} X + Y &= \{x + y \mid x \in X \wedge y \in Y\} \\ X - Y &= \{x - y \mid x \in X \wedge y \in Y\} \\ X * Y &= \{x * y \mid x \in X \wedge y \in Y\}. \end{aligned}$$

There is, however, no consensus on interval division.

In most expositions, interval division, X/Y, is only defined under the condition that 0 not be contained in Y. Such a restriction is acknowledged by several authors to be unacceptable and unnecessary. However, for a long time the obvious remedy in Definition 4 below has been considered an exotic variant of interval arithmetic with two inaccessible publications [Kahan 1968; Hanson 1968] as sole references. Only recently [Novos 1993; Ratz 1996; Walster 1998] has interval division received the attention it needed.

In Novos [1993] and Ratz [1996], the intervals X and Y were required to be bounded intervals. As a result, these systems were not closed under division.

We define the quotient X/Y of two intervals as follows:

Definition 4. Let X and Y be real intervals, then we define

$$X/Y = \{z \mid \exists x \in X, y \in Y \text{ such that } y \neq 0,\ z = x/y\}$$

Observe that this defines the quotient of two intervals to be a set which may not itself be an interval. For example,

$$\{1\}/\{x \mid x \leq 1\} = \{x \mid x < 0\} \cup \{x \mid 1 \leq x\}.$$

One of the main contributions of this paper is to provide explicit formulas for this quotient when X and Y are real intervals (Theorem 8 in Section 4.7) and more generally when they are only assumed to be connected sets of reals (Theorem 16 in Section 6).

It is well known that the bounded intervals are closed under the arithmetic operations, provided one disallows division by intervals containing zero. We state and prove this theorem below for completeness. In the next section, we compute explicit formulas for these operations on unbounded intervals and in this way prove that the addition, subtraction, and multiplication operations are closed on possibly unbounded intervals, whereas division is not.

THEOREM 2. *If S and T are nonempty, bounded, real intervals, then so are $S + T$, $S - T$, and $S * T$. If, in addition, T does not contain zero, then S/T is a nonempty, bounded, real interval as well. More generally, $f(S, T)$ will be a bounded, real interval, provided f is continuous on a domain containing $S \times T$.*

PROOF. Note that according to Definition 2, "real interval" means *closed* connected set. This theorem is a consequence of some of the general properties of continuous functions, namely, that continuous functions map connected sets into connected sets, and map compact sets into compact sets. Since the compact sets of $\mathcal{R}$ are just the closed and bounded subsets, the theorem follows from the fact that the first three operators are continuous on $\mathcal{R}^2$ and the division operator is continuous on $\mathcal{R} \times (\mathcal{R} \setminus \{0\})$. □

Of course, $\{x/y \mid x \in X \wedge y \in Y \wedge y \neq 0\}$ is unbounded if $X = \{u \in \mathcal{R} \mid 0 \leq u \leq 1\}$ and if $Y = \{u \in \mathcal{R} \mid 0 < u \leq 1\}$. But such a Y is not an interval, because it is not closed. If we take a closed set for Y, such as $\{u \in \mathcal{R} \mid 0 < \epsilon \leq u \leq 1\}$, then, as the theorem says, X/Y is closed and bounded, hence an interval.

In the case where T contains zero or is unbounded, the quotient S/T may not be an interval. Indeed, even though $S = \{1\}$ and $T = \{x \in \mathcal{R} \mid -1 \leq x \leq 1\}$ are intervals and

$$S/T = \{x \in \mathcal{R} \mid (x \leq -1) \vee (1 \leq x)\}$$

is defined as a set of reals, S/T is not a connected set, hence not an interval. Similarly, although $S = \{1\}$ and $T = \{x \in \mathcal{R} \mid x \geq 1\}$ are intervals, the set $S/T = \{x \in \mathcal{R} \mid 0 < x \leq 1\}$ is not closed, and hence not an interval.

3. *Interval Evaluation of Expressions*

Correctness in interval arithmetic means that interval evaluation of any expression has to yield an interval containing all possible values in terms of reals. We need to make such a requirement precise, which is what we do in this section.

3.1. AN INCLUSION PROPERTY. Any form of interval arithmetic should have a correctness property that may be formulated as follows. Let $v_1, \ldots, v_n$ be the variables that occur in an expression E. Let p be the real that results from evaluating E with real values $a_1, \ldots, a_n$ substituted for $v_1, \ldots, v_n$. Let I be the result of evaluating E with intervals $I_1, \ldots, I_n$ substituted for the variables $v_1, \ldots, v_n$ where $a_1 \in I_1, \ldots, a_n \in I_n$. The property then requires that $p \in I$.

Here we show that the inclusion property for our approach follows from a more general lemma in set theory. There is no need to restrict the inclusion property to sets in the form of intervals, or to total functions. In view of division, it is essential not to restrict it to total functions. We restrict our consideration to real functions only for notational simplicity—it is easy to see that these results can be extended to many-sorted functions over arbitrary domains.

Moore [1966, Theorem 3.1] was the first to prove what was independently dubbed by Hansen [1992] and Rall [1969] to be "The Fundamental Theorem of

Interval Arithmetic." Alefeld and Herzberger [1983] formulate such a property for continuous real functions (presumably total functions). Walster and Hansen [1998] generalized the theorem to apply to functions on intervals of reals that do not need to have the inclusion isotonicity property and that do not need to be interval extensions.

In connection with the inclusion property, it is important to distinguish between expressions and the functions computed by these expressions. Thus, before we review the required set-theoretic matters, we first consider expressions and their evaluation.

3.2. EXPRESSIONS. The constituents of an expression are variables, function symbols, parentheses and commas. An expression is recursively defined as being a variable or as being $f(e_1, \ldots, e_n)$ where f is a function symbol and $e_1, \ldots, e_n$ are expressions, with $n \geq 0$. If $n = 0$, then $f()$ can be regarded as a constant.

Values of expressions depend on the values of their variables. This is formalized by means of *environments* mapping variables to values. Thus, an expression E may have a value in a given environment. When E has the value y in environment $\mathcal{V}$, we write $y = E|\mathcal{V}$. Here, we take $\mathcal{V}$ to be an assignment of a value $a = \mathcal{V}(v)$ to each variable v and a partial function $\phi = \mathcal{V}(f)$ to each function symbol f. Because the partial functions may not be total, the environment may fail to assign a value to the expression. We sometimes write $\mathcal{V}(v \mapsto a,\ f \mapsto \phi)$ for an environment mapping v to a and f to ϕ.

We now define recursively the result of evaluating an expression. If the expression E is a variable v, then $E|\mathcal{V}(v \mapsto a)$ is a. If E is a compound expression $f(e_1, \ldots, e_n)$ where the e_i are expressions, then we have

$$f(e_1, \ldots, e_n)|\mathcal{V} = \mathcal{V}(f)(e_1|\mathcal{V}, \ldots, e_n|\mathcal{V}).$$

Thus, if the expressions $e_1, \ldots, e_n$ have values $a_1, \ldots, a_n$ under $\mathcal{V}$, and if $\phi = \mathcal{V}(f)$, then

$$E|\mathcal{V} = \phi(a_1, \ldots, a_n).$$

The generality afforded by partial functions is important because, for example, division is not a total function.

Let us illustrate these definitions by a treatment of the expression x/y. We consider an environment $\mathcal{V}$ such that $\mathcal{V}(/)$ is the real division operator, then $x/y|\mathcal{V}(x \mapsto 1, y \mapsto 1) = 1$, while $x/y|\mathcal{V}(x \mapsto 1, y \mapsto 0)$ is undefined.

3.3. SOME PREREQUISITES FROM SET THEORY. Let $f : S \rightarrow T$ be a partial function. We call S the *source* and T the *target*. The domain of f, $dom(f)$, is the subset of S on which f is defined. The range of f, $ran(f)$, is $\{f(x) \mid x \in dom(f)\}$.

One may wonder why to distinguish between domain and source? Why not make them equal by definition, as is done in some treatments? The reason is that to be able to compose functions, as is needed in the evaluation of nested expressions, the target of one function needs to equal the source of another. Thus, for a language of expressions to be useful, one standardizes on as few sets as possible for sources and targets. Here the choice of standardization is obvious: the set of reals for all sources and targets. Yet we want to include division among the functions, so it is useful if domains can differ from sources.

Definition 5. Let $f : S \to T$ be any partial function and let 2^S (respectively 2^T) denote the set of all subsets of S (respectively T). Then, a function $F : 2^S \to 2^T$ is said to be a *set extension* of f if for all $X \subset S$ we have

$$\forall x \in X \cap dom(f).\, f(x) \in F(X).$$

Definition 6. Let $f : S \to T$ be any partial function and let $\hat{f} : 2^S \to 2^T$ be defined by $\forall X \subset S$,

$$\hat{f}(X) = \{ f(x) \mid x \in X \cap dom(f)\}.$$

Here $\hat{f}$ is the *canonical set extension* of f.

LEMMA 1. *Let $f : S \to T$ be any partial function, then $\hat{f}$ is the smallest set extension of f, that is, $\hat{f}$ is a set extension of f and if F is any set extension, then one must have*

$$\forall X \subset S.\ \hat{f}(X) \subset F(X).$$

Moreover, $\hat{f}$ is a monotone function in the sense that for all subsets U, V of S, we have

$$U \subset V \ \Rightarrow\ \hat{f}(U) \subset \hat{f}(V).$$

Note that this lemma holds for all *partial* functions f. It is clear from the definition that the canonical set extension is *total*, that is, is defined on *all* subsets of S. Translated to interval arithmetic this implies that if intervals are regarded as sets of reals, and the interval operations are defined as canonical set extensions, which is what we do, then a *closed system* results.

Note that the above also applies if the source S of $f : S \to T$ is a Cartesian product $S_1 \times \cdots \times S_n$. Such a function has n arguments and can be assigned by an environment to an n-place function symbol.

We are now ready to formulate the general set-theoretic lemma of which properties like the "Fundamental Theorem of Interval Arithmetic" are instances.

LEMMA 2. *Let E be an expression, and $\mathcal{V}$ an environment for E, and let $\tilde{\mathcal{V}}$ be an environment that assigns a set $\tilde{\mathcal{V}}(v)$ to the variable v such that $\mathcal{V}(v) \in \tilde{\mathcal{V}}(v)$. Let $\tilde{\mathcal{V}}$ assign to every function symbol f a function $\tilde{\mathcal{V}}(f)$ that is a set extension of $\mathcal{V}(f)$. Then,*

—*$E \mid \tilde{\mathcal{V}}$ exists (even though $E \mid \mathcal{V}$ may not),*
—*If $E \mid \mathcal{V}$ exists, then $E \mid \mathcal{V} \in E \mid \tilde{\mathcal{V}}$.*

Thus, when we evaluate an expression with canonical set extensions as interpretation for the function symbols, we obtain a set that contains all values it should contain according to the inclusion property.

4. *Syntactical Treatment of Real Intervals and Their Operations*

Although our semantics has defined real intervals and their operations in a mathematically rigorous way, so far we could only use cumbersome set-comprehension expressions such as

$$\{x \in \mathcal{R} \mid a \leq x \leq b\}.$$

What we need in addition are concise expressions for real intervals. We also need rules for computing the operations of Definitions 3 and 4 on the basis of such expressions. These expressions and their manipulation we regard as the *syntactical* aspect of interval arithmetic.

4.1. Expressions for Real Intervals

Definition 7. Let a and b be reals such that $a \leq b$.

$$\begin{aligned}\langle a, b\rangle &\overset{\text{def}}{=} \{x \in \mathcal{R} \mid a \leq x \leq b\}\\ \langle -\infty, b\rangle &\overset{\text{def}}{=} \{x \in \mathcal{R} \mid x \leq b\}\\ \langle a, +\infty\rangle &\overset{\text{def}}{=} \{x \in \mathcal{R} \mid a \leq x\}\\ \langle -\infty, +\infty\rangle &\overset{\text{def}}{=} \mathcal{R}\end{aligned}$$

The definition gives an expression for each of the types of nonempty real interval that exist according to Theorem 1. To take full advantage of the notation, we regard each expression abstractly as a *pair*. The first (second) element of a pair is called left (right) endpoint of the interval denoted by the pair.

Thus, we can summarize all expressions of the definition by $\langle a, b\rangle$ where a and b belong to the set $\mathcal{R} \cup \{-\infty, +\infty\}$ of *extended reals* and $a \leq b$. The above notations do not cover the empty interval. We have not found it urgent to find a special notation for it and will use $\emptyset$.

We have chosen to use an angle bracket notation $\langle a, b\rangle$ to denote these real intervals so as to avoid any confusion with the square bracket notation $[a, b]$ used for "extended real intervals" by other authors (e.g., Walster [1998]). In the latter, the notation $[a, b]$ defines a set S of extended reals which always contains both of its endpoints a, b. Our notation with angle brackets always denotes sets of reals and the endpoint is contained in the set if and only if the endpoint is a real. Thus, with our notation an infinite endpoint (which is not a real) implies that the set is unbounded on that side. Our angle bracket notation is a syntax that employs the extended reals (possibly infinite) for specifying sets of reals (each of which, by its nature, is finite, though a set of them may not be bounded).

Below, we summarize the properties of the extended reals. We note that

COROLLARY 1. *If $\langle a, b\rangle$ is a nonempty real interval, then $a \neq +\infty$ and $b \neq -\infty$.*

PROOF. See Definition 7 and Theorem 1 that says that there are no other nonempty intervals than the ones covered by Definition 7. □

This corollary will turn out to be useful for avoiding undefined operations.

4.2. EXAMPLES. Now that we have a convenient notation for intervals, let us illustrate by means of examples some of the consequences of our semantic definitions of the interval operations.

(1) $\langle 2, 2\rangle * \langle \pi, \pi\rangle = \langle 2\pi, 2\pi\rangle$. In other approaches this holds because intervals are generalized reals. In our approach this is true because in Definition 3 all reals in the set $\{2\}$ combine with all reals in the set $\{\pi\}$ to produce $\{2\pi\}$.

(2) $\langle 0, 0\rangle * \langle -\infty, \infty\rangle = \langle 0, 0\rangle$. This is easily verified with Definition 3.

(3) $\langle 0, 1\rangle/\langle 0, 1\rangle = \langle 0, \infty\rangle$. This holds as all nonnegative numbers, but only those, can be expressed as x/y with $x, y \in \langle 0, 1\rangle$.

(4) $\langle 1, 1\rangle/\langle -\infty, +\infty\rangle = \mathcal{R} \setminus \{0\}$. This is not a real interval. In fact, it is not even connected.

(5) $\langle 1, 1\rangle/\langle -1, 1\rangle = \langle -\infty, -1\rangle \cup \langle +1, +\infty\rangle$. This is a disjoint union of two real intervals.

(6) $\langle 1, 1\rangle/\langle -1, \infty\rangle = \langle -\infty, -1\rangle \cup \langle 0, +\infty\rangle \setminus \{0\}$. This is neither closed nor connected.

Although addition, subtraction, and multiplication of nonempty intervals always produces nonempty intervals, the same is not true for division, as shown by the following theorem.

THEOREM 3. *Let S and T be nonempty real intervals, then S/T is empty if and only if $T = \langle 0, 0\rangle$.*

PROOF. Observe that if T contains a nonzero element, then S/T is nonempty. Hence, if S/T is empty, then T can contain only 0, so we must have $T = \langle 0, 0\rangle$. Conversely, $S/\langle 0, 0\rangle = \emptyset$ by Definition 4. □

4.3. CLASSIFICATION OF NONEMPTY, REAL INTERVALS ACCORDING TO SIGN. Later, it will be useful to distinguish several cases for interval multiplication and division according to the signs of the elements of real intervals. It turns out to be sufficient to distinguish according to whether an interval contains a positive number, and within each of the resulting subclasses, to distinguish according to whether the interval contains a negative number. Thus, there are four cases to consider.

—The class M (“Mixed”) is defined as the set of real intervals containing at least one positive and at least one negative real. Thus, for all intervals $\langle a, b\rangle$ in class M, we have $a < 0 < b$.

—The class Z (“Zero”) is defined as the set of nonempty real intervals containing neither a positive nor a negative number. $Z = \{\langle 0, 0\rangle\}$.

—The class P (“Positive”) is defined as the set of real intervals containing at least one positive, but no negative number. It follows that $0 \leq a \leq b$ and $0 < b$. We further partition P into class P_0 (those intervals for which $a = 0$) and P_1 (where $a > 0$). Note that because P_0 contains at least one positive element, $\langle a, b\rangle \in P_0$ implies that $b > 0$. Hence, $\langle a, b\rangle \in P$ implies that $b > 0$.

—The class N (“Negative”) is the symmetric counterpart of P: $a \leq b \leq 0$ and $a < 0$. We further partition N into class N_0 (those intervals for which $b = 0$) and N_1 (where $b < 0$). Note that because N_0 contains at least one negative element, $\langle a, b\rangle \in N_0$ implies that $a < 0$. Hence, $\langle a, b\rangle \in N$ implies that $a < 0$.

We use the classification $\{M, P, Z, N\}$ to define interval multiplication. The further partitioning of P by $\{P_0, P_1\}$ and of N by $\{N_0, N_1\}$ is needed for interval division.

Our classification is summarized in the table in Figure 1.

4.4. SUMMARY OF THE EXTENDED REALS. The extended reals are the set $\mathcal{R} \cup \{-\infty, +\infty\}$. Extended reals a and b are ordered as among the reals, if both are real. Moreover, $-\infty < c < +\infty$ for any real c.

Class of $\langle a, b\rangle$	at least one negative	at least one positive	Signs of endpoints
M	yes	yes	$a < 0 \wedge b > 0$
Z	no	no	$a = 0 \wedge b = 0$
P	no	yes	$a \geq 0 \wedge b > 0$
P_0	no	yes	$a = 0 \wedge b > 0$
P_1	no	yes	$a > 0 \wedge b > 0$
N	yes	no	$a < 0 \wedge b \leq 0$
N_0	yes	no	$a < 0 \wedge b = 0$
N_1	yes	no	$a < 0 \wedge b < 0$

FIG. 1. Classification of nonempty intervals by sign. As only nonempty intervals are classified, we have $a \leq b$.

				x		
	$x + y$	$-\infty$	NR	0	PR	$+\infty$
	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$\perp$
	NR		NR	NR	$\mathcal{R}$	$+\infty$
y	0			0	PR	$+\infty$
	PR				PR	$+\infty$
	$+\infty$					$+\infty$

				x		
	$x - y$	$-\infty$	NR	0	PR	$+\infty$
	$-\infty$	$\perp$	$+\infty$	$+\infty$	$+\infty$	$+\infty$
	NR	$-\infty$	$\mathcal{R}$	PR	PR	$+\infty$
y	0	$-\infty$	NR	0	PR	$+\infty$
	PR	$-\infty$	NR	NR	$\mathcal{R}$	$+\infty$
	$+\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$\perp$

				x		
	$x * y$	$-\infty$	NR	0	PR	$+\infty$
	$-\infty$	$+\infty$	$+\infty$	$\perp$	$-\infty$	$-\infty$
	NR		PR	0	NR	$-\infty$
y	0			0	0	$\perp$
	PR				PR	$+\infty$
	$+\infty$					$+\infty$

				x		
	x/y	$-\infty$	NR	0	PR	$+\infty$
	$-\infty$	$\perp$	0	0	0	$\perp$
	NR	$+\infty$	PR	0	NR	$-\infty$
y	0	$\perp$	$\perp$	$\perp$	$\perp$	$\perp$
	PR	$-\infty$	NR	0	PR	$+\infty$
	$+\infty$	$\perp$	0	0	0	$\perp$

FIG. 2. The arithmetic operations on the extended reals. Omitted entries are defined by symmetry. The $\perp$ symbol indicates an undefined operation. PR indicates a positive real; NR indicates a negative real.

The arithmetic operations on $\mathcal{R}$ are extended to the extended reals as specified in Figure 2. The symbol $\perp$ indicates a case where the operation is not defined.

4.5. INTERVAL ADDITION AND SUBTRACTION. If $\langle a, b\rangle$ and $\langle c, d\rangle$ are bounded, nonempty, real intervals, then Theorem 2 guarantees that $\langle a, b\rangle + \langle c, d\rangle$, $\langle a, b\rangle - \langle c, d\rangle$, and $\langle a, b\rangle * \langle c, d\rangle$ are bounded, nonempty, real intervals as well. In this section, we derive rules for the expressions for these result intervals with endpoints obtained by extended-real operations on a, b, c, and d, and we verify that these rules hold for unbounded intervals as well.

THEOREM 4. *If $\langle a, b\rangle$ and $\langle c, d\rangle$ are real intervals, then*

$$\begin{aligned} \langle a, b\rangle + \langle c, d\rangle &= \langle a + c, b + d\rangle, \text{ and} \\ \langle a, b\rangle - \langle c, d\rangle &= \langle a - d, b - c\rangle. \end{aligned}$$

Moreover, $a + c$, $b + d$, $a - d$, and $b - c$ are defined as extended reals.

PROOF. The expressions for the intervals follow from Definition 3 of interval addition and subtraction and from the monotonicity of addition and subtraction. That they hold when a and/or c is $-\infty$ and when b and/or d is $+\infty$ is easily checked, for example, if b and/or d is infinite, then the sum contains arbitrarily large elements and so must have right endpoint ∞. The fact that the undefined expressions $+\infty + (-\infty)$, $-\infty + (+\infty)$, $+\infty - (+\infty)$, and $-\infty - (-\infty)$ cannot

Class of $\langle a,b\rangle$	Class of $\langle c,d\rangle$	Left Endpoint of $\langle a,b\rangle * \langle c,d\rangle$	Right Endpoint of $\langle a,b\rangle * \langle c,d\rangle$	Symmetry
P	P	$a*c$	$b*d$	proved directly
P	M	$b*c$	$b*d$	proved directly
P	N	$b*c$	$a*d$	$x*y=-(x*-y)$
M	P	$a*d$	$b*d$	$x*y=y*x$
M	M	$\min(a*d,b*c)$	$\max(a*c,b*d)$	proved directly
M	N	$b*c$	$a*c$	$x*y=-(x*-y)$
N	P	$a*d$	$b*c$	$x*y=-(-x*y)$
N	M	$a*d$	$a*c$	$x*y=-(-x*y)$
N	N	$b*d$	$a*c$	$x*y=-(x*-y)$
Z	P,M,N,Z	0	0	proved directly
P,M,N	Z	0	0	proved directly

FIG. 3. Case analysis for multiplication of real intervals, $\langle a, b\rangle * \langle c, d\rangle$.

arise in $a + c$, $b + d$, $a - d$, or $b - c$ follows from $a \neq +\infty$, $b \neq -\infty$, $c \neq +\infty$, and $d \neq -\infty$, in accordance with Corollary 1. □

4.6. INTERVAL MULTIPLICATION. In the formulas, for interval addition it was sufficient to ensure that the expressions for the endpoints are defined in such a way that a and c cannot be $+\infty$ and that b and d cannot be $-\infty$. In the case of multiplication the undefined case is $0 * \pm\infty$. Our classification scheme (see the table in Figure 1) is designed to help avoid these cases.

THEOREM 5. *If $\langle a, b\rangle$ and $\langle c, d\rangle$ are bounded, real intervals, then*

$$\langle a, b\rangle * \langle c, d\rangle = \langle \min(S), \max(S)\rangle,$$

where $S = \{a * c, a * d, b * c, b * d\}$.

PROOF. This is a well-known and easy-to-prove result (see e.g., Moore [1996]). We include the proof here for completeness. Since $\langle a, b\rangle$ and $\langle c, d\rangle$ are closed, bounded sets of reals, a, b, c, and d must all be real numbers. By Definition 3, the set $\langle a, b\rangle * \langle c, d\rangle$, is equal to

$$\{x * y \mid x \in \langle a, b\rangle \wedge y \in \langle c, d\rangle\}.$$

Since $x * y$ is continuous in both arguments, this set is bounded from below and contains its greatest lower bound. As $x * y$ has no local minimum in $\langle a, b\rangle \times \langle c, d\rangle$ (nor indeed anywhere in $\mathcal{R} \times \mathcal{R}$), the greatest lower bound must occur at one of the four corners of $\langle a, b\rangle \times \langle c, d\rangle$. A similar reasoning shows that $\langle a, b\rangle \times \langle c, d\rangle$ contains its least upper bound, and that this is also in S. □

Observe that this theorem does not immediately extend to the case of unbounded intervals. Indeed, if a, b, c, d are allowed to be infinite, then the products $a * c$, $a * d$, $b * c$, $b * d$ may have the form $0 * \pm\infty$ and so the terms $\min(S)$ and $\max(S)$ are no longer defined. These problems with unbounded intervals are resolved in the following theorem by decomposing the problem into nine subproblems based on the classification in the table in Figure 1. This decomposition has the added benefit of reducing the number of products one must compute from eight to zero, two, or four, depending on the classification (See Figure 6).

THEOREM 6. *If $\langle a, b\rangle$ and $\langle c, d\rangle$ are real intervals, then $\langle a, b\rangle * \langle c, d\rangle$ is a real interval whose endpoints are given by the expressions in Figure* 3.

PROOF. We only need to calculate the endpoints in three of the nine cases. The remaining cases can then be obtained via some of the symmetries of $x * y$.

Let us first consider the case MM, that is, $\langle a, b\rangle \in M$ and $\langle c, d\rangle \in M$, where M is the set of intervals defined in the table in Figure 1. We split the intervals into subintervals over which multiplication is monotonic.

$$\begin{aligned}\langle a, b\rangle * \langle c, d\rangle &= \langle a, 0\rangle * \langle c, 0\rangle \cup \langle a, 0\rangle * \langle 0, d\rangle \cup \\ &\quad\ \langle 0, b\rangle * \langle c, 0\rangle \cup \langle 0, b\rangle * \langle 0, d\rangle \\ &= \langle 0, a * c\rangle \cup \langle a * d, 0\rangle \cup \langle b * c, 0\rangle \cup \langle 0, b * d\rangle \\ &= \langle \min(a * d, b * c), \max(a * c, b * d)\rangle.\end{aligned}$$

In case MM, none of the endpoints a, b, c, or d is zero. Hence, none of $a * d$, $a * c$, $b * c$, or $b * d$ can be undefined as extended real.

Next, we consider the case PP. Here, $x * y$ is monotonic in both arguments. Hence, $\langle a, b\rangle * \langle c, d\rangle = \langle a * c, b * d\rangle$. In case PP, the possibilities for zero and infinity are segregated in the expression $\langle a*c, b*d\rangle$: zeros can only occur in the left endpoint; infinities only in the right endpoint. Thus, neither expression can become undefined as an extended real.

One more case needs to be considered. Let us take PM.

$$\begin{aligned}\langle a, b\rangle * \langle c, d\rangle &= \langle a, b\rangle * \langle c, 0\rangle \cup \langle a, b\rangle * \langle 0, d\rangle \\ &= \langle b * c, 0\rangle \cup \langle 0, b * d\rangle = \langle b * c, b * d\rangle.\end{aligned}$$

In case PM, a can be zero, but b, c, d cannot. As a does not occur in either endpoint, neither endpoint can become undefined as an extended real.

The remaining cases can be obtained by applying symmetry. For example, we can use $x * y = y * x$. That is, the case MP is obtained from PM by interchanging a and c and interchanging b and d.

Another useful symmetry is based on the identity $x * y = -(-x * y)$. This is realized by first interchanging a and b (this takes care of the inner minus sign) and interchanging in the result the right and left endpoints (for the outer minus sign). This symmetry gives NM from PM and NP from PP.

The table in Figure 3 can be completed by one further symmetry: the one based on $x * y = -(x * -y)$, which is implemented by interchanging first c and d and interchanging in the result the right and left endpoints. This gives PN from PP, MN from MP, and NN from NP. □

4.7. INTERVAL DIVISION. Interval computation involving division has to be a compromise between information gain, computational efficiency, and program complexity. For example, as a set, the interval quotient of $\langle 1, 1\rangle$ and $\langle -\infty, 1\rangle$ is

$$\{x/y \mid\ x \in \langle 1, 1\rangle, y \in \langle -\infty, 1\rangle, y \neq 0\}$$

and this simplifies to

$$\{x/y \mid\ x = 1, y \leq 1, y \neq 0\} = \{x \mid x < 0\} \cup \{x \mid 1 \leq x\}.$$

For efficiency in computation, one may choose to represent sets of reals by a single closed interval. This can be simply achieved by replacing $\langle a, b\rangle/\langle c, d\rangle$ by the least interval containing it. For the example above, this would yield the interval $\langle -\infty, +\infty\rangle$. Another choice would be to represent this quotient as a finite union of closed intervals, which would result in more information but also a greater cost in storage and processing for the operation. For the example above, this would yield $\langle -\infty, 0\rangle \cup \langle 1, +\infty\rangle$, which captures all information except for the openness of the right endpoint of the first interval. The final extreme would represent the quotient as a finite union of intervals together with endpoint data (as we consider in Section 6). This maintains all information about the quotient but at a considerable cost in program complexity. Indeed, the computation of the quotient of two intervals with possibly open endpoints is a fairly complex operation as we will see in Theorem 16 of Section 6. In this section, we are only concerned with determining all the facts about $\langle a, b\rangle / \langle c, d\rangle$, regardless of what a software designer judges to be worth using.

Interval multiplication is simple in that the result is always an interval, and this interval is characterized by the formula in Theorem 6. The only work in addition to this was to identify all possible ways in which this formula can be optimized and to verify that results of operations are always defined. In interval division, all this needs to be done also. In addition to that, we have to deal with the complication that $\langle a, b\rangle / \langle c, d\rangle$ might not be an interval.

In the case where both numerator and denominator are bounded, real intervals, Ratz [1996] has provided a formula for computing an interval quotient $X \oslash Y$ when X and Y are bounded intervals, which may possibly contain zero and he has proved that his formula is correct. His definition of interval division is somewhat different from the one we propose. For the purposes of this paper, we call Ratz's division (as defined in Definition 8 below) the *relational division* operator, and we refer to ours (as defined in Definition 4 above) as the *functional division* operator.

Definition 8. The relational division operator $\oslash$ is defined by

$$\langle a, b\rangle \oslash \langle c, d\rangle = \{z \in \mathcal{R} \mid \exists x, y.\ a \le x \le b, c \le y \le d, x = y * z\}$$

Given this definition, the Ratz [1996] formula is given by the following theorem:

THEOREM 7 [RATZ 1996]. *Let $\langle a, b\rangle$ and $\langle c, d\rangle$ be two nonempty bounded real intervals. Then $\langle a, b\rangle \oslash \langle c, d\rangle =$*

$$\begin{cases}
\langle a, b\rangle * \langle 1/d, 1/c\rangle & \textit{if } 0 \notin \langle c, d\rangle \\
\langle -\infty, \infty\rangle & \textit{if } 0 \in \langle a, b\rangle \ \wedge\ 0 \in \langle c, d\rangle \\
\langle b/c, \infty\rangle & \textit{if } b < 0 \ \wedge c < d = 0 \\
\langle -\infty, b/d\rangle \cup \langle b/c, \infty\rangle & \textit{if } b < 0 \ \wedge c < 0 < d \\
\langle -\infty, b/d\rangle & \textit{if } b < 0 \ \wedge 0 = c < d \\
\langle -\infty, a/c\rangle & \textit{if } 0 < a \ \wedge c < d = 0 \\
\langle -\infty, a/c\rangle \cup \langle a/d, \infty\rangle & \textit{if } 0 < a \ \wedge c < 0 < d \\
\langle a/d, \infty\rangle & \textit{if } 0 < a \ \wedge 0 = c < d \\
\emptyset & \textit{if } 0 \notin \langle a, b\rangle \ \wedge\ c = d = 0.
\end{cases}$$

Ratz proves this by considering each of these cases and deriving the result by a series of direct transformations from the definition of $X \oslash Y$.

We first observe that the relational division definition extends naturally to real intervals as we define them, even though Ratz only defines them for closed and bounded intervals. In this more general context, we see that our division is more general in the sense that relational division can easily be calculated from our division, as shown in the following lemma:

LEMMA 3. *Let X and Y be real intervals, then $X/Y \subset X \oslash Y$ and*

$$X \oslash Y = \begin{cases} X/Y & \text{if } 0 \notin X \cap Y \\ \mathcal{R} & \text{otherwise.} \end{cases}$$

PROOF. First, observe that if $x/y = z$, then $x = y * z$, so $X/Y \subset X \oslash Y$. The converse is true if 0 is not in both X and Y. Indeed, if $0 \notin Y$, then $x = y * z$ implies that $y \neq 0$ and $z = x/y$. Similarly, if $0 \notin X$, then $x = y * z$ implies $y \neq 0$ and so $z = x/y$. On the other hand, if 0 is contained in both X and Y, then $X \oslash Y = \mathcal{R}$ as $0 = 0 * z$ holds for all real z. □

There are several reasons for extending Ratz's theorem.

(1) It only defines division among bounded intervals. The theorem shows that the quotient of two bounded intervals is either empty, or a bounded interval, or an unbounded interval, or a union of two unbounded intervals. Although the result can be an unbounded interval (or a union of two unbounded intervals), the Ratz formula does not allow the arguments to be unbounded intervals. Of course, Ratz intended this extended interval division only to be used in the context of the Interval Newton method, where the possibly unbounded set would be intersected with the original bounded interval, to give zero, one, or two bounded intervals. It turns out to be hardly more complex to define a general-purpose interval division.

(2) Theorem 7 relies on the multiplication formulas by converting many of the quotients into products $\langle a, b\rangle * \langle 1/d, 1/c\rangle$. This can be inefficient and also can introduce additional roundoff errors (as, for example, a/d will in general be more precise than $a * (1/d)$ when evaluated in floating point arithmetic).

(3) It only computes the result of the relational division $X \oslash Y$, but there are times when functional division is more appropriate. For example, if we evaluate $xy/(x^2 + y^2)$ on the interval $x = y = \langle 0, 1\rangle$ using functional division we obtain $\langle 0, 1\rangle/\langle 0, 2\rangle = \langle 0, \infty\rangle$ which shows that the function is nonnegative on that interval, whereas using relational division yields $\langle 0, 1\rangle \oslash \langle 0, 2\rangle = \langle -\infty, \infty\rangle$ which conveys no information.

Of course, if one allows division by unbounded intervals one admits a complication that Ratz did not have to handle: the resulting interval is no longer guaranteed to be a closed set. For example, $\langle 1, 1\rangle/\langle 1, \infty\rangle = \{x \in \mathcal{R} \mid 0 < x \leq 1\}$ is a connected set that does not contain its greatest lower bound, and hence is not a closed set. The following theorem shows that this complication is conveniently handled by our classification of intervals.

THEOREM 8. *If $\langle a, b\rangle$ and $\langle c, d\rangle$ are nonempty, real intervals, then their functional quotient can be computed as follows. If either is $\langle 0, 0\rangle$, then*

$$\langle a, b\rangle/\langle 0, 0\rangle = \emptyset, \qquad \langle 0, 0\rangle/\langle c, d\rangle = \begin{cases} \langle 0, 0\rangle & \text{if } \langle c, d\rangle \neq \langle 0, 0\rangle \\ \emptyset & \text{if } \langle c, d\rangle = \langle 0, 0\rangle \end{cases}$$

If neither is equal to $\langle 0, 0\rangle$, then $\langle a, b\rangle/\langle c, d\rangle$ is given as in the "general formula" column of the table in Figure 4, unless the specified condition in column 4 holds, in which case the quotient is given by the exception case formula in column 5.

Class of $\langle a,b\rangle$	Class of $\langle c,d\rangle$	$\langle a,b\rangle/\langle c,d\rangle$ general formula	unless	$\langle a,b\rangle/\langle c,d\rangle$ exception case	
P_1	P	$\langle a/d, b/c\rangle \setminus \{0\}$	$c=0$	$\langle a/d, \infty\rangle \setminus \{0\}$	D
P_0	P	$\langle 0, b/c\rangle$	$c=0$	$\langle 0, \infty\rangle$	D
M	P	$\langle a/c, b/c\rangle$	$c=0$	$\langle -\infty, \infty\rangle$	D
N_0	P	$\langle a/c, 0\rangle$	$c=0$	$\langle -\infty, 0\rangle$	S_2
N_1	P	$\langle a/c, b/d\rangle \setminus \{0\}$	$c=0$	$\langle -\infty, b/d\rangle \setminus \{0\}$	S_2
P_1	M	$(\langle -\infty, a/c\rangle \cup \langle a/d, \infty\rangle) \setminus \{0\}$			D
P_0	M	$\langle -\infty, +\infty\rangle$			D
M	M	$\langle -\infty, +\infty\rangle$			D
N_0	M	$\langle -\infty, +\infty\rangle$			S_2
N_1	M	$(\langle -\infty, b/d\rangle \cup \langle b/c, \infty\rangle) \setminus \{0\}$			S_2
P_1	N	$\langle b/d, a/c\rangle \setminus \{0\}$	$d=0$	$\langle -\infty, a/c\rangle \setminus \{0\}$	S_1
P_0	N	$\langle b/d, 0\rangle$	$d=0$	$\langle -\infty, 0\rangle$	S_1
M	N	$\langle b/d, a/d\rangle$	$d=0$	$\langle -\infty, \infty\rangle$	S_1
N_0	N	$\langle 0, a/d\rangle$	$d=0$	$\langle 0, \infty\rangle$	S_2
N_1	N	$\langle b/c, a/d\rangle \setminus \{0\}$	$d=0$	$\langle b/c, \infty\rangle \setminus \{0\}$	S_2

FIG. 4. Case analysis for functional division of real intervals, $\langle a, b\rangle/\langle c, d\rangle$ when $a \le b$, $c \le d$, and neither interval is $\langle 0, 0\rangle$. The last column refers to how the formula has been proved ("D" for a direct proof, "S_1" and "S_2" refer to a symmetry used to reduce it to an earlier case.) The "class" labels, N, N_1, N_0, M, P_0, P_1, P are as in Figure 1.

PROOF. Before beginning the proof, we make the observation that the exception cases in the table all arise because the general formula contains a quotient of the form u/v with $u \neq 0$, and in the exception case $v = 0$. We see in the next section that the IEEE signed zero properties can be used to entirely eliminate the exception column, thereby greatly simplifying the computation of interval quotients.

In accordance with the table in Figure 4, we prove the cases MM (i.e., $\langle a, b\rangle \in M$ and $\langle c, d\rangle \in M$), P_0M, P_0P, P_1P, MP, and P_1M directly. These six directly proved cases are indicated by a D in the last column. Then, we use the fact that N is the symmetrical counterpart of P in accordance with symmetry $x/y = -(x/-y)$ (indicated by S_1 in column six). This gives MN from MP, P_0N from P_0P, and P_1N from P_1P. Finally, we obtain all six cases where $\langle a, b\rangle \in N_1$ or $\langle a, b\rangle \in N_0$ from those where $\langle a, b\rangle \in P_1$ or $\langle a, b\rangle \in P_0$ by using symmetry $x/y = -(-x/y)$, (indicated by S_2). Thus, it remains to prove table entries MM, P_0M, P_0P, P_1P, MP, P_1M.

In cases MM and P_0M, we have $\langle 0, +\epsilon\rangle \subset \langle a, b\rangle$ and $\langle -\epsilon, +\epsilon\rangle \subset \langle c, d\rangle$ for some $\epsilon > 0$. This ensures that all reals occur in $\langle a, b\rangle/\langle c, d\rangle$, so that the quotient interval is $\langle -\infty, +\infty\rangle$.

Case P_0P: $a = 0 < b$ and $0 \le c$. If $c = 0$, then the quotient contains $\langle 0, \epsilon\rangle/\langle 0, \epsilon\rangle = \langle 0, \infty\rangle$ and contains no negative values so the exception case must be $\langle 0, \infty\rangle$. If $c \neq 0$, then $c > 0$ and so $\langle 0, b\rangle/\langle c, d\rangle = \langle 0, b/c\rangle$. This holds also if b or d is $+\infty$.

Case P_1P: $0 < a$ and $0 \le c$. Note that a and c are finite. Suppose first that $c \neq 0$, then $\langle a, b\rangle$ and $\langle c, d\rangle$ are single signed and nonzero, so $\langle a, b\rangle/\langle c, d\rangle = \langle a/d, b/c\rangle$, provided that b and d are finite. The formula holds when $b = +\infty$ because the quotient contains arbitrarily large numbers, so the right endpoint should be ∞. If d is infinite, then $a/d = a/\infty = 0$ by the rules of extended-real arithmetic, but since $0 \notin \langle a, b\rangle/\langle c, d\rangle$ we include the possibility of unbounded $\langle c, d\rangle$ by excluding 0: $\langle a, b\rangle/\langle c, d\rangle = \langle a/d, b/c\rangle \setminus \{0\}$. If $c = 0$, then the quotient contains arbitrarily large positive values, so the exception case is $\langle a/d, \infty\rangle \setminus \{0\}$.

Let us next consider case *MP*: $a < 0 < b$ and $0 \leq c$. If c is zero, then the quotient $\langle a, b\rangle / \langle c, d\rangle$ contains $\langle -\epsilon, \epsilon\rangle / \langle 0, \epsilon\rangle$ for some $\epsilon > 0$, so the result should be $\langle -\infty, \infty\rangle$ in this exception case. Otherwise, $c > 0$, so splitting into single signed components and using our formula for case P_0P we get

$$\begin{aligned}\langle a, b\rangle / \langle c, d\rangle &= \langle a, 0\rangle / \langle c, d\rangle \cup \langle 0, b\rangle / \langle c, d\rangle \\ &= \langle a/c, 0\rangle \cup \langle 0, b/c\rangle \\ &= \langle a/c, b/c\rangle\end{aligned}$$

This formula also holds for $a = -\infty$ and/or $b = +\infty$ as c is finite and so $\pm\infty / c = \pm\infty$.

Case P_1M: $0 < a$ and $c < 0 < d$. Note that a is finite. Splitting into single signed components, and using our formula for case P_1P, we get

$$\begin{aligned}\langle a, b\rangle / \langle c, d\rangle &= \langle a, b\rangle / \langle c, 0\rangle \cup \langle a, b\rangle / \langle 0, d\rangle \\ &= (\langle -\infty, a/c\rangle \setminus \{0\}) \cup (\langle a/d, +\infty\rangle \setminus \{0\}) \\ &= (\langle -\infty, a/c\rangle \cup \langle a/d, +\infty\rangle) \setminus \{0\}\end{aligned}$$

The right-hand side is a union of connected sets that are closed except when one of the endpoints is zero. □

For completeness, we also provide the formulas for the extension of Ratz' relational division.

COROLLARY 2. *If $\langle a, b\rangle$ and $\langle c, d\rangle$ are nonempty, real intervals, then their relational quotient can be computed as follows. If either is $\langle 0, 0\rangle$, then*

$$\langle a, b\rangle \oslash \langle 0, 0\rangle = \begin{cases} \emptyset & \text{if } 0 \notin \langle a, b\rangle \\ \mathcal{R} & \text{if } 0 \in \langle a, b\rangle \end{cases}$$

$$\langle 0, 0\rangle \oslash \langle c, d\rangle = \begin{cases} \langle 0, 0\rangle & \text{if } 0 \notin \langle c, d\rangle \\ \mathcal{R} & \text{if } 0 \in \langle c, d\rangle \end{cases}$$

If neither is $\langle 0, 0\rangle$, then $\langle a, b\rangle \oslash \langle c, d\rangle$ is given by the functional division table in Figure 4, except that the results in the exception column for the four cases N_0N, N_0P, P_0N, P_0P should be replaced by $\langle -\infty, \infty\rangle$.

PROOF. By Theorem 3, we know that X/Y and $X \oslash Y$ are equal except possibly in the case where both numerator and denominator contain zero, in which case $X \oslash Y$ is $\langle -\infty, \infty\rangle$, while X/Y can also be one of $\langle 0, \infty\rangle$, $\langle -\infty, 0\rangle$, $\langle 0, 0\rangle$, or $\emptyset$. Thus, the cases of division by $\langle 0, 0\rangle$ and of $\langle 0, 0\rangle$ being divided, must be modified to check for the case when the other interval contains zero. Also one must potentially modify all other cases where 0 can be in both numerator and denominator. These consist of the four cases mentioned above N_0N, N_0P, P_0N, P_0P, as well as N_0M, MN_0, P_0M, MP_0, MM. These last five yield $\langle -\infty, \infty\rangle$ for functional division also, so the only places the table must be changed is the four cases specified in the corollary. □

4.8. THE INTERVAL NEWTON METHOD. One important application of interval arithmetic is the Interval Newton method, which can be regarded as an interval analog of the classical version of Newton's method for finding a zero of a function. Both methods make use of the first-order version of Taylor's theorem, which states that for any function $f : \mathcal{R} \to \mathcal{R}$, which is continuously differentiable on an interval I, one must have:

$$\forall a, x \in I, \exists \xi \in I \text{ such that } \; f(x) = f(a) + (x - a) f'(\xi).$$

The Interval Newton method attempts to use information about $f(a)$ and the range of f' on I to contract the set of possible zeroes of f within I.

In this section, we show that the Interval Newton method can be implemented using the interval arithmetic described in this paper and that this method applies even when the interval I is unbounded. The following two theorems provide the main properties of the Interval Newton method and the proofs are along the same lines as those given in Hansen [1992].

THEOREM 9. *Let f be a function that is continuously differentiable in an interval I and let F and F' be any set extensions of f and f', let $a \in I$ and let $N_a = a - (F(a)/F'(I))$. If $0 \notin F(a) \cap F'(I)$, then all zeroes of f in I are contained in $I \cap N_a$.*

PROOF. If $f(x) = 0$, then we have, for some $\xi \in I$, $0 = f(a) + f'(\xi)(x - a)$. If $0 \notin F'(I)$, then $f'(\xi) \neq 0$ and so $x = a - f(a)/f'(\xi)$. Similarly, if $0 \notin F(a)$, we have that $f(a) \neq 0$, which implies that $f'(\xi)(x - a) \neq 0$, and again $f'(\xi) \neq 0$. It follows that $x = a - f(a)/f'(\xi)$ for some $\xi \in I$. Since F and F' are set extensions of f and f', we must have $f(a) \in F(a)$ and $f'(\xi) \in F'(I)$ and hence x is in the interval $a - (F(a)/F'(I))$ as claimed. □

THEOREM 10. *Notation as in Theorem* 9. *If*

(*a*) $N_a \subset I$,

(*b*) N_a *is nonempty and bounded*, *and*

(*c*) $0 \notin F'(I)$,

then f has a unique zero in N_a.

PROOF. Observe that we have

$$\{a - (f(a)/f'(\xi)) : \; \xi \in I\} \subset N_a$$

and hence if $N_a = \langle u, v\rangle$ for $u, v \in I \cap \mathcal{R}$, then

$$u \leq a - f(a)/f'(\xi) \leq v$$

for all $\xi \in I$. Since we are assuming $0 \notin F'(I)$ and f' is continuous on I, we know that either $f'(\xi) > 0$ for all $\xi \in I$ or $f'(\xi) < 0$ for all $\xi \in I$. In the first case, this implies that

$$\begin{aligned} f(a) + f'(\xi_1)(u - a) \; &\leq \; 0 \;\; \forall \xi_1 \in I \\ f(a) + f'(\xi_2)(v - a) \; &\geq \; 0 \;\; \forall \xi_2 \in I, \end{aligned}$$

hence, $f(u) \leq 0 \leq f(v)$, and by continuity of f, it must have a zero in I. If $f'(I) < 0$, then we infer that $f(u) \geq 0 \geq f(v)$ and again that f has a zero in I. Finally, the assumption that $0 \notin F'(I)$ implies that f is strictly monotone in I and hence it must have a unique zero. □

Theorems 9 and 10 provide the basis for a root finding algorithm very similar to that described by Hansen [1992]. This algorithm begins with a (possibly unbounded) interval I and an expression E which represents a function f that is continuously differentiable on I. The algorithm uses the interval arithmetic operations defined in this paper to compute set extensions F and F' of f and f' (respectively). It then proceeds by selecting an element $a \in I$ and contracting I to $I \cap N_a$ if $0 \notin F(a)$ or $0 \notin F'(I)$. If this condition does not hold or if $I \subset N_a$, then the interval I can be split into two smaller intervals and the algorithm is recursively applied. The recursion is terminated when intervals are sufficiently small or when splitting has produced too many intervals. This algorithm returns a set of intervals with the property that their union contains all zeros, if any, of the function.

Observe that there is certainty about the absence of zeros outside the intervals returned by this algorithm. Whether there actually exist any zeros inside any given interval, and if so, whether there is exactly one, can often be determined by applying the test in Theorem 10.

Note that we could follow Ratz [1996] and use relational division $\oslash$ rather than functional division in the Interval Newton iteration As the following theorem (and its proof) makes clear, this use of relational division makes the application of the Interval Newton method more uniform, but it does *not* result in more contraction or in an enhanced ability to detect the existence of zeroes and their uniqueness.

THEOREM 11. *Notation as in Theorem* 9, *and let*

$$\widehat{N_a} = a - (F(a) \oslash F'(I))$$

(*i*) *All zeroes of f in I are contained in $\widehat{N_a} \cap I$,*

(*ii*) *If*

(*a*) *$\widehat{N_a} \subset I$, and*

(*b*) *$\widehat{N_a}$ is nonempty and bounded,*

then f has a unique zero in $\widehat{N_a}$.

PROOF. The only difference between functional and interval division occurs when both numerator and demoninator contain zero, and in this case $\widehat{N_a} = \langle -\infty, +\infty \rangle$. So (i) follows by the result of Theorem 9 with the observation that if $0 \in F(a) \cap F'(I)$, then $\widehat{N_a} \cap I = I$ and so the conclusion is trivially true. Observe that this does not yield a better contraction, it just makes the contraction somewhat simpler to state and provides less information about when the contraction is likely to be useful. Similarly, to prove (ii) we observe that the condition that $0 \notin F'(I)$ could be dropped from Theorem 10 because $0 \in F'(I)$ implies that $\widehat{N_a}$ is either empty (if $F'(I) = \langle 0, 0 \rangle$ and $0 \notin F(a)$) or unbounded (if $0 \in F(a)$ or $F'(I) \neq \langle 0, 0 \rangle$), and so doesn't meet the hypotheses of the Theorem. Note that using relational division simplifies the test for whether there is a unique zero (by making one test redundant), but it computes exactly the same result as when functional division is used. □

5. *Exploiting the IEEE-Standard for Interval Arithmetic*

So far, our considerations have only taken into account the properties of the reals and the extended reals. Our method for representing sets of reals by pairs of extended reals has the property that endpoints of computed intervals are defined according to the arithmetic of extended reals even when infinity is involved. In the next section, we first give an outline of the IEEE 754 standard for floating point arithmetic. Then we show how it can be used to ensure that endpoint computation *always* yields a defined, correct, and optimal result. In other words, from the point of view of interval arithmetic, the standard extends the extended reals in just the right way.

5.1. OVERVIEW OF THE IEEE 754 STANDARD. In this section, we review the IEEE-standard floating-point number system as far as needed for this paper. The standard specifies several formats differing only in the sizes of certain fields. In this article, we are only concerned with features of the standard common to all formats.

For any particular format, the set of possible bit patterns is partitioned into the following categories:

(1) nonzero reals

(2) -0, $+0$, $-\infty$, and $+\infty$

(3) bit patterns that do not represent reals and are called NaN (Not a Number)

The IEEE standard orders the non-NaN floating-point numbers of a given format as follows: $-\infty$, the negative real floating-point numbers in increasing order, -0, $+0$, the positive real floating-point numbers in increasing order, $+\infty$.

In this article, we consider the operation of addition, subtraction, multiplication, and division. The standard specifies a resulting floating-point number for each operation on each of the bit patterns, whether or not a corresponding mathematical definition exists. There is a mathematical definition, according to the field of reals, only if both operands are reals (therefore, not if either is -0, $+0$, $-\infty$, or $+\infty$). If a mathematical definition applies, then the resulting real may not be a floating-point number. In such cases, the standard specifies that the result is one of the endpoints of the least interval of reals with non-NaNs as endpoints that contains the result according to the field of reals. For the purpose of this sentence, $-\infty < x < -0$ for all negative real x, $+0 < x < +\infty$ for all positive real x, and $-0 = 0 = +0$. Which of the two endpoints is selected as result depends on the *rounding mode* selected in the operation of the floating-point number system. In this article, we consider the mode of *downward rounding*, where the lesser endpoint is selected, and the mode of *upward rounding*, where the greater endpoint is selected.

In cases where the field of reals does not provide a result, the standard specifies as result the one of a limiting process, if one is unambiguously suggested by the operands. In other cases, $+0 + (-0)$ and $+0 - (+0)$, the result is arbitrarily defined. In the remaining cases, the result is a NaN. These considerations are summarized in the tables of Figure 5.

In this article, we are only concerned with some of the operations specified in the standard. For each of these, we will need the result rounded in a specific direction. Thus, we rely on combinations of the operations $+$, $-$, $*$, $/$.

Definition 9. The operations of addition, subtraction, multiplication, and division of the IEEE standard with rounding towards $-\infty$ are denoted $+_{lo}$, $-_{lo}$, $*_{lo}$, and $/_{lo}$ respectively. The same operations with rounding towards $+\infty$ are denoted $+_{hi}$, $-_{hi}$, $*_{hi}$, and $/_{hi}$ respectively.

The standard also requires each operation to set a Boolean flag *exact* which will be true if and only if the computed result is equal to the mathematically defined result. In this case, no rounding takes place, so the rounding mode is irrelevant.

5.2. THE SIGNED ZERO CONVENTION. In this section, we show how signed zeroes can be used to simplify the formulas for interval addition and multiplication.

				x			
	$x + y$	$-\infty$	NR	-0	$+0$	PR	$+\infty$
	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	NaN
	NR		FR'	FR	FR	FR	$+\infty$
y	-0			-0	±0	FR	$+\infty$
	$+0$				$+0$	FR	$+\infty$
	PR					FR'	$+\infty$
	$+\infty$						$+\infty$

				x			
	$x - y$	$-\infty$	NR	-0	$+0$	PR	$+\infty$
	$-\infty$	NaN	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$
	NR	$-\infty$	FR	FR	FR	FR'	$+\infty$
y	-0	$-\infty$	FR	±0	$+0$	FR	$+\infty$
	$+0$	$-\infty$	FR	-0	±0	FR	$+\infty$
	PR	$-\infty$	FR'	FR	FR	FR	$+\infty$
	$+\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	NaN

				x			
	$x * y$	$-\infty$	NR	-0	$+0$	PR	$+\infty$
	$-\infty$	$+\infty$	$+\infty$	NaN	NaN	$-\infty$	$-\infty$
	NR		FR'	$+0$	-0	FR'	$-\infty$
y	-0			$+0$	-0	-0	NaN
	$+0$				$+0$	$+0$	NaN
	PR					FR'	$+\infty$
	$+\infty$						$+\infty$

				x			
	x/y	$-\infty$	NR	-0	$+0$	PR	$+\infty$
	$-\infty$	NaN	$+0$	$+0$	-0	-0	NaN
	NR	$+\infty$	FR'	$+0$	-0	FR'	$-\infty$
y	-0	$+\infty$	$+\infty$	NaN	NaN	$-\infty$	$-\infty$
	$+0$	$-\infty$	$-\infty$	NaN	NaN	$+\infty$	$+\infty$
	PR	$-\infty$	FR'	-0	$+0$	FR'	$+\infty$
	$+\infty$	NaN	-0	-0	$+0$	$+0$	NaN

FIG. 5. The arithmetic operations on the IEEE Standard floating-point numbers. The *FR*, *FR*′ entries denote a result obtained according to the mathematical definition of the field of reals and then rounded according to the selected or default rounding mode. Such rounding results in a non-NaN floating-point number. The result in some cases is finite (*FR*); in others it may be finite or infinite (*FR*′). In the addition/subtraction tables, ±0 is +0 in all rounding modes except when the rounding mode towards $-\infty$, and then it is -0. *NR* stands for negative real; *PR* stands for positive real.

In particular, we use the fact that if we let -0 and $+0$ denote the signed zeroes of IEEE arithmetic, then division by signed zero $x/(-0)$ and $x/(+0)$ is a non-NAN floating point number for all non-zero x. If we introduce signed zeroes into our extended real arithmetic, all of the exception formulas that arose in Figure 4 are properly handled by the general formulas, provided we adopt the convention that all zero endpoints are signed zeroes and $+0$ is used for left endpoints, while -0 is used for right endpoints. For example, consider the last line of Figure 4, which gives the formula when $b < 0$ and $d \leq 0$:

$$\langle a, b\rangle/\langle c, d\rangle = \langle b/c, a/d\rangle \setminus \{0\}$$

unless $d = 0$, in which case

$$\langle a, b\rangle/\langle c, d\rangle = \langle b/c, \infty\rangle \setminus \{0\}$$

If we adopt the convention that (-0) is used when d is zero, then since a is negative and nonzero, the IEEE specification on division of signed zeroes implies that $a/d = +\infty$, and so the exception case is not needed. The following definition of IEEE intervals adopts this convention:

Definition 10. An *IEEE-standard interval* is a real interval whose endpoints are represented by IEEE floating point numbers. We further require that -0 can only appear as a right endpoint, and $+0$ can only appear as a left endpoint.

With this convention, we find that the IEEE standard facilitates interval arithmetic to a remarkable extent (e.g., compare Figures 9 and 11 below, the latter uses signed zeroes). However, we realize that the beauty of the standard is in the eye of the beholder: other researchers in interval arithmetic [Stolfi and de Figueiredo 1997] criticize the signed zeros as follows:

> While it is possible to concoct examples where this feature saves an instruction or two, in the vast majority of applications this value is an annoying distraction and a source of subtle bugs.

Our convention removes any ambiguity about what sign of zero should appear in which endpoint. Walster [1998] goes further and assigns different meanings to ± 0 and $\pm\infty$ depending on whether they appear in the left or right endpoint. These extra zeroes are used to represent underflow values that are between zero and the smallest positive floating point value. Similarly, the extra infinities are used to represent overflow values and are operationally similar to our use of $\pm\infty$ in the endpoints of an interval.

5.3. Optimal IEEE Approximations of Interval Arithmetic. In interval arithmetic, rounding need not lead to error. By rounding outward, correctness is maintained and rounding only has the effect of including some values that would have been left out were the result exact. The following is an obvious fact. The reason for presenting it as a theorem is its great importance.

Theorem 12. *For every set of reals, there is a unique narrowest IEEE-standard floating-point interval containing it.*

Proof. Included among the IEEE-standard floating-point numbers are $-\infty$ and $+\infty$. Hence, there exists such an interval containing the given set of reals. As the number of such intervals is finite and closed under intersection, there is a least interval containing the given set. □

One reason for the great importance of this theorem is that the uniqueness of the existing least containing interval compels the definition of the following function:

Definition 11. For any set α of reals, $\Gamma(\alpha)$ is the least floating-point interval containing it.

Definition 12. We call a set α a *sound approximation* of a set β of reals if $\alpha \supseteq \beta$. We call the IEEE interval $\Gamma(\beta)$ the *optimal IEEE approximation* of β.

For addition, subtraction, and multiplication, it is a simple matter to obtain sound and optimal approximations:

Theorem 13. *Let $X = \langle a, b\rangle$ and $Y = \langle c, d\rangle$ be nonempty IEEE-standard intervals, then*

$$\begin{aligned}\Gamma(\langle a, b\rangle + \langle c, d\rangle) &= \langle a +_{lo} c, b +_{hi} d\rangle \text{ and}\\ \Gamma(\langle a, b\rangle - \langle c, d\rangle) &= \langle a -_{lo} d, b -_{hi} c\rangle\end{aligned}$$

The formulas in Figures 6 and 7 give sound approximations to $X * Y$ and X/Y respectively. The former give the optimal IEEE approximation of $X * Y$. The latter is contained in an optimal IEEE approximation of X/Y, but is more informative.

Proof. First, note that X/Y may not be connected, and hence, when it contains two components, we compute the optimal approximation of each component. Moreover, X/Y may not be closed and our tables also indicate this occurrence by using the set difference operator $A \setminus \{0\}$.

The formulas in the Theorem and in Figures 6 and 7 are obtained from the corresponding formulas in the case of real intervals (Figures 3 and 4) by using outward rounding, that is, rounding right endpoints toward positive infinity and left endpoints toward negative infinity. It is clear that this results in a sound approximation. Optimality follows from the fact that the upward rounded arithmetic operations is required by the IEEE standard to return the smallest floating point number which is

Class of $\langle a,b\rangle$	Class of $\langle c,d\rangle$	$\Gamma(\langle a,b\rangle * \langle c,d\rangle)$
P	P	$\langle a*_{lo}c, b*_{hi}d\rangle$
M	P	$\langle a*_{lo}d, b*_{hi}d\rangle$
N	P	$\langle a*_{lo}d, b*_{hi}c\rangle$
P	M	$\langle b*_{lo}c, b*_{hi}d\rangle$
M	M	$\langle \min(a*_{lo}d, b*_{lo}c), \max(b*_{hi}d, a*_{hi}c)\rangle$
N	M	$\langle a*_{lo}d, a*_{hi}c\rangle$
P	N	$\langle b*_{lo}c, a*_{hi}d\rangle$
M	N	$\langle b*_{lo}c, a*_{hi}c\rangle$
N	N	$\langle b*_{lo}d, a*_{hi}c\rangle$
Z	P, M, N	$\langle 0, 0\rangle$
P, M, N, Z	Z	$\langle 0, 0\rangle$
any	$\emptyset$	$\emptyset$
$\emptyset$	any	$\emptyset$

FIG. 6. Multiplication of IEEE intervals.

Class of $\langle a,b\rangle$	Class of $\langle c,d\rangle$	a sound approximation of $\langle a,b\rangle/\langle c,d\rangle$	$\Gamma(\langle a,b\rangle/\langle c,d\rangle)$
P_1	P	$\langle a/_{lo}d, b/_{hi}c\rangle \setminus \{0\}$	$\langle a/_{lo}d, b/_{hi}c\rangle$
P_0	P	$\langle 0, b/_{hi}c\rangle$	$\langle 0, b/_{hi}c\rangle$
M	P	$\langle a/_{lo}c, b/_{hi}c\rangle$	$\langle a/_{lo}c, b/_{hi}c\rangle$
N_0	P	$\langle a/_{lo}c, 0\rangle$	$\langle a/_{lo}c, 0\rangle$
N_1	P	$\langle a/_{lo}c, b/_{hi}d\rangle \setminus \{0\}$	$\langle a/_{lo}c, b/_{hi}d\rangle$
P_1	M	$(\langle -\infty, a/_{hi}c\rangle \cup \langle a/_{lo}d, +\infty\rangle) \setminus \{0\}$	$\langle -\infty, +\infty\rangle$
P_0, M, N_0	M	$\langle -\infty, +\infty\rangle$	$\langle -\infty, +\infty\rangle$
N_1	M	$(\langle -\infty, b/_{hi}d\rangle \cup \langle b/_{lo}c, +\infty\rangle) \setminus \{0\}$	$\langle -\infty, +\infty\rangle$
P_1	N	$\langle b/_{lo}d, a/_{hi}c\rangle \setminus \{0\}$	$\langle b/_{lo}d, a/_{hi}c\rangle$
P_0	N	$\langle b/_{lo}d, 0\rangle$	$\langle b/_{lo}d, 0\rangle$
M	N	$\langle b/_{lo}d, a/_{hi}d\rangle$	$\langle b/_{lo}d, a/_{hi}d\rangle$
N_0	N	$\langle 0, a/_{hi}d\rangle$	$\langle 0, a/_{hi}d\rangle$
N_1	N	$\langle b/_{lo}c, a/_{hi}d\rangle \setminus \{0\}$	$\langle b/_{lo}c, a/_{hi}d\rangle$
Z	P, M, N	$\langle 0, 0\rangle$	$\langle 0, 0\rangle$
P, M, N, Z	Z	$\emptyset$	$\emptyset$
any	$\emptyset$	$\emptyset$	$\emptyset$
$\emptyset$	any	$\emptyset$	$\emptyset$

FIG. 7. Functional division of IEEE intervals. These formulas require that $\langle c, d\rangle$ adheres to the signed zero convention of Section 5.2, that is, $c \neq -0,\ d \neq +0$.

not smaller than the true result, and similarly for downward rounded operations. We are also using the signed zero convention, which eliminates the exception conditions of Theorem 8. □

6. *Arithmetic on Connected Subsets of the Reals*

The results of the previous sections extend to the more general class of connected subsets of $\mathcal{R}$.

Definition 13. A *general real interval* is a connected set of reals.

To represent such an interval X syntactically, we must provide both its endpoints $\langle a, b\rangle$ and two bits of information α and β, with α (respectively, β) specifying whether a (respectively, b) belongs to X. We formalize this in the following definition:

Definition 14. Let $\mathcal{R}^* = \mathcal{R} \cup \{-\infty, \infty\}$ denote the extended reals. For any $u, v \in \mathcal{R}^*$ and any Boolean values $\alpha, \beta \in \mathbf{B} = \{t, f\}$, let $\langle u, v\rangle_{\alpha,\beta}$ denote the set X of all real values between u and v, where α (respectively, β) is true iff u (respectively, v) is contained in X. Thus,

$$\begin{aligned}
\langle u, v\rangle_{t,t} &= \{x \in R \mid u \leq x \leq v\} \\
\langle u, v\rangle_{t,f} &= \{x \in R \mid u \leq x < v\} \\
\langle u, v\rangle_{f,t} &= \{x \in R \mid u < x \leq v\} \\
\langle u, v\rangle_{f,f} &= \{x \in R \mid u < x < v\}.
\end{aligned}$$

Note that this only defines subsets of the reals. A consequence of this definition is that, for example, $\langle u, v\rangle_{t,f} = \langle u, v\rangle_{t,t}$ if $v = +\infty$. Of course, $\langle u, v\rangle_{t,f} \neq \langle u, v\rangle_{t,t}$ whenever $u, v \in \mathcal{R}$ and $u < v$.

In our interval arithmetic formulas for general real intervals, we use intersections and unions of general intervals. We include the relevant formulas below.

THEOREM 14. *Let $X = \langle a, b\rangle_{\alpha,\beta}$ and $Y = \langle c, d\rangle_{\gamma,\delta}$ be general real intervals and suppose $X \cap Y \neq \emptyset$. Then,*

$$\begin{aligned}
\langle a, b\rangle_{\alpha,\beta} &\cap \langle c, d\rangle_{\gamma,\delta} \\
&= \langle \max(a, c), \min(b, d)\rangle_{\mu(a,c,\gamma,\alpha),\mu(b,d,\beta,\delta)} \\
\langle a, b\rangle_{\alpha,\beta} &\cup \langle c, d\rangle_{\gamma,\delta} \\
&= \langle \min(a, c), \max(b, d)\rangle_{\nu(a,c,\alpha,\gamma),\nu(b,d,\delta,\beta)}
\end{aligned}$$

where

$$\begin{aligned}
\mu(a, c, \alpha, \gamma) &= ((a < c) \wedge \alpha) \vee ((a = c) \wedge (\alpha \wedge \gamma)) \\
&\quad \vee ((a > c) \wedge \gamma) \\
\nu(a, c, \alpha, \gamma) &= ((a < c) \wedge \alpha) \vee ((a = c) \wedge (\alpha \vee \gamma)) \\
&\quad \vee ((a > c) \wedge \gamma)
\end{aligned}$$

PROOF. The only subtle point here is to determine whether or not each endpoint is contained in the result interval. For an intersection, the left endpoint $L = \max(a, c)$ is either a or c, whichever is larger. If a is larger, then L is contained in the intersection if and only if a is in X. Similarly, if c is the larger, L is in the intersection if and only if c is in Y. If a and c are equal, then L is in the intersection if and only if both $a \in X$ and $c \in Y$; similarly, with the right endpoint. For the union of two general intervals, the similar arguments apply except that when $a = c$, L is in the union if and only if $a \in X$ or $c \in Y$. □

We now provide the formulas for interval arithmetic on general real intervals.

THEOREM 15. *Let $X = \langle a, b\rangle_{\alpha,\beta}$ and $Y = \langle c, d\rangle_{\gamma,\delta}$ be general real intervals. Then,*

$$\begin{aligned}
\langle a, b\rangle_{\alpha,\beta} + \langle c, d\rangle_{\gamma,\delta} &= \langle a + c, b + d\rangle_{\alpha\wedge\gamma,\beta\wedge\delta} \\
\langle a, b\rangle_{\alpha,\beta} - \langle c, d\rangle_{\gamma,\delta} &= \langle a - d, b - c\rangle_{\alpha\wedge\delta,\beta\wedge\gamma}
\end{aligned}$$

Class of $\langle a,b\rangle_{\alpha,\beta}$	Class of $\langle c,d\rangle_{\gamma,\delta}$	$\langle a,b\rangle_{\alpha,\beta} * \langle c,d\rangle_{\gamma,\delta}$
P	P	$\langle a*c, b*d\rangle_{\psi(a,c,\alpha,\gamma),\psi(b,d,\beta,\delta)}$
P	M	$\langle b*c, b*d\rangle_{\psi(b,c,\beta,\gamma),\psi(b,d,\beta,\delta)}$
P	N	$\langle b*c, a*d\rangle_{\psi(b,c,\beta,\gamma),\psi(a,d,\alpha,\delta)}$
M	P	$\langle a*d, b*d\rangle_{\psi(a,d,\alpha,\delta),\psi(b,d,\beta,\delta)}$
M	M	$\langle a*d, b*d\rangle_{\psi(a,d,\alpha,\delta),\psi(b,d,\beta,\delta)} \cup \langle b*c, a*c\rangle_{\psi(b,c,\beta,\gamma),\psi(a,c,\alpha,\gamma)}$
M	N	$\langle b*c, a*c\rangle_{\psi(b,c,\beta,\gamma),\psi(a,c,\alpha,\gamma)}$
N	P	$\langle a*d, b*c\rangle_{\psi(a,d,\alpha,\delta),\psi(b,c,\beta,\gamma)}$
N	M	$\langle a*d, a*c\rangle_{\psi(a,d,\alpha,\delta),\psi(a,c,\alpha,\gamma)}$
N	N	$\langle b*d, a*c\rangle_{\psi(b,d,\beta,\delta),\psi(a,c,\alpha,\gamma)}$
Z	P, M, N	$\langle 0,0\rangle_{t,t}$
P, M, N, Z	Z	$\langle 0,0\rangle_{t,t}$
any	$\emptyset$	$\emptyset$
$\emptyset$	any	$\emptyset$

FIG. 8. Multiplication of general real intervals where $\psi(a, c, \alpha, \gamma) = (\alpha \wedge \gamma) \vee (\alpha \wedge (a = 0)) \vee (\gamma \wedge (c = 0))$.

Class of $\langle a,b\rangle_{\alpha,\beta}$	Class of $\langle c,d\rangle_{\gamma,\delta}$	$\langle a,b\rangle_{\alpha,\beta} / \langle c,d\rangle_{\gamma,\delta}$
P	P_1	$\langle a/d, b/c\rangle_{\psi(a,d,\alpha,\delta),\psi(b,c,\beta,\gamma)}$
P	P_0	$\langle a/d, \infty\rangle_{\psi(a,d,\alpha,\delta),f}$
P	M	$\langle -\infty, a/c\rangle_{f,\psi(a,c,\alpha,\gamma)} \cup \langle a/d, \infty\rangle_{\psi(a,d,\alpha,\delta),f}$
P	N_0	$\langle -\infty, a/c\rangle_{f,\psi(a,c,\alpha,\gamma)}$
P	N_1	$\langle b/d, a/c\rangle_{\psi(b,d,\beta,\delta),\psi(a,c,\alpha,\gamma)}$
M	P_1	$\langle a/c, b/c\rangle_{\psi(a,c,\alpha,\gamma),\psi(b,c,\beta,\gamma)}$
M	P_0, M, N_0	$\langle -\infty, +\infty\rangle_{f,f}$
M	N_1	$\langle b/d, a/d\rangle_{\psi(b,d,\beta,\delta),\psi(a,d,\alpha,\delta)}$
N	P_1	$\langle a/c, b/d\rangle_{\psi(a,c,\alpha,\gamma),\psi(b,d,\beta,\delta)}$
N	P_0	$\langle -\infty, b/d\rangle_{f,\psi(b,d,\beta,\delta)}$
N	M	$\langle -\infty, b/d\rangle_{f,\psi(b,d,\beta,\delta)} \cup \langle b/c, \infty\rangle_{\psi(b,c,\beta,\gamma),f}$
N	N_0	$\langle b/c, \infty\rangle_{\psi(b,c,\beta,\gamma),f}$
N	N_1	$\langle b/c, a/d\rangle_{\psi(b,c,\beta,\gamma),\psi(a,d,\alpha,\delta)}$
Z	P, M, N	$\langle 0,0\rangle_{t,t}$
any	$Z, \emptyset$	$\emptyset$
$\emptyset$	any	$\emptyset$

FIG. 9. Functional division of general real intervals where $\psi(a, c, \alpha, \gamma) = (\alpha \wedge \gamma) \vee (\alpha \wedge (a = 0)) \vee (\gamma \wedge (c = 0))$.

and $X * Y$, and X/Y are given by the formulas in the tables in Figures 8 *and* 9. *These tables make use of the Boolean function*

$$\psi(a, c, \alpha, \gamma) = (\alpha \wedge \gamma) \vee (\alpha \wedge (a = 0)) \vee (\gamma \wedge (c = 0)).$$

PROOF. The multiplication formulas in Figure 8 are identical to the formulas in Figure 3 for the multiplication of closed intervals except that the endpoints of the result interval in the closed case may or may not be contained in the set in this more general case. For the case of nonzero endpoints, it is easy to see that the endpoint $(u * v)$ is contained in the set if and only if the corresponding endpoints (u and v) appearing in the formula for that endpoint are contained in their sets. On the other hand, zero endpoints only appear in the result of an interval multiplication if one or both of the argument intervals has a zero endpoint, and clearly the result contains zero precisely if zero is contained in either of the argument intervals. These observations are captured by the Boolean function ψ, given in the theorem, which is used in the tables. Observe in particular that

$$\begin{aligned}(a = 0) \wedge (c = 0) \;\Rightarrow\; \psi(a, c, \alpha, \gamma) \;&=\; \alpha \vee \gamma\\ (a = 0) \wedge (c \neq 0) \;\Rightarrow\; \psi(a, c, \alpha, \gamma) \;&=\; \alpha\\ (a \neq 0) \wedge (c = 0) \;\Rightarrow\; \psi(a, c, \alpha, \gamma) \;&=\; \gamma\\ (a \neq 0) \wedge (c \neq 0) \;\Rightarrow\; \psi(a, c, \alpha, \gamma) \;&=\; \alpha \wedge \gamma.\end{aligned}$$

For interval division, the formulas in Figure 9 are obtained from the formulas in Figure 4 by combining a few observations. First, consider the case where $\langle a, b\rangle$ and $\langle c, d\rangle$ are both in P. From Figure 4, we have

$X = \langle a, b\rangle$	$Y = \langle c, d\rangle$	X/Y
P_1	P_1	$\langle a/d, b/c\rangle \setminus \{0\}$
P_1	P_0	$\langle a/d, \infty\rangle \setminus \{0\}$
P_0	P_1	$\langle a/d, b/c\rangle$
P_0	P_0	$\langle a/d, \infty\rangle$.

We must remove $\{0\}$ from the quotient in the first two cases because it may happen that $d = \infty$ in which case $a/d = 0$, but since 0 cannot be expressed as x/y with (finite) real numbers $x \in X$ and $y \in Y$, it is not in the quotient.

Consider now the case where X and Y are general intervals of type P, then it is not hard to see that a/d is contained in the quotient if and only if $(a \in X) \wedge (d \in Y)$ or $(a = 0) \wedge (a \in X)$. In particular, in the case where $d = \infty$, we must have $d \notin Y$ as Y is a set of real numbers, and hence $0 = a/d \notin X/Y$. Thus, this rule automatically handles the removal of extraneous zeroes from the quotient. For the right endpoint b/c, we see that it is in the quotient if and only if both b and c are. Hence, the general case of $X, Y \in P$, where $X = \langle a, b\rangle_{\alpha,\beta}$, $Y = \langle c, d\rangle_{\gamma,\delta}$ can be summarized by the following rule

$$X/Y = \begin{cases} \langle a/d, b/c\rangle_{\psi(a,d,\alpha,\delta),\psi(b,c,\beta,\gamma)} & \text{if } c > 0\\ \langle a/d, \infty\rangle_{\psi(a,d,\alpha,\delta),\, f} & \text{if } c = 0,\end{cases}$$

where the fact that $b, c, d \neq 0$ implies that

$$\begin{aligned}\psi(a, d, \alpha, \delta) \;&=\; (\alpha \wedge \delta) \vee ((a = 0) \wedge \alpha)\\ \psi(b, c, \beta, \gamma) \;&=\; (\beta \wedge \gamma).\end{aligned}$$

This explains the P/P_1 and P/P_0 rows in Figure 9. The other cases P/N,N/P,N/N,M/P,M/N,M/M,P/M,N/M are obtained by similar arguments, and result in the formulas in Figure 9. □

To obtain optimal approximations for the multiplication and division of (nonclosed) intervals, we need to define optimal approximation in the context of general intervals.

Definition 15. The *optimal general IEEE approximation* of a set U is a set S satisfying

—$U \subset S$,

—S is a finite union of general intervals with IEEE endpoints, and

—no smaller such S exists.

We denote S, if it exists, by $\Gamma^*(U)$.

Class of $\langle a,b\rangle_{\alpha,\beta}$	Class of $\langle c,d\rangle_{\gamma,\delta}$	$\Gamma^*(\langle a,b\rangle_{\alpha,\beta} * \langle c,d\rangle_{\gamma,\delta})$
P	P	$\langle a*_{lo}c, b*_{hi}d\rangle_{\psi'(a,c,\alpha,\gamma),\psi'(b,d,\beta,\delta)}$
P	M	$\langle b*_{lo}c, b*_{hi}d\rangle_{\psi'(b,c,\beta,\gamma),\psi'(b,d,\beta,\delta)}$
P	N	$\langle b*_{lo}c, a*_{hi}d\rangle_{\psi'(b,c,\beta,\gamma),\psi'(a,d,\alpha,\delta)}$
M	P	$\langle a*_{lo}d, b*_{hi}d\rangle_{\psi'(a,d,\alpha,\delta),\psi'(b,d,\beta,\delta)}$
M	M	$\langle a*_{lo}d, b*_{hi}d\rangle_{\psi'(a,d,\alpha,\delta),\psi'(b,d,\beta,\delta)} \cup \langle b*_{lo}c, a*_{hi}c\rangle_{\psi'(b,c,\beta,\gamma),\psi'(a,c,\alpha,\gamma)}$
M	N	$\langle b*_{lo}c, a*_{hi}c\rangle_{\psi'(b,c,\beta,\gamma),\psi'(a,c,\alpha,\gamma)}$
N	P	$\langle a*_{lo}d, b*_{hi}c\rangle_{\psi'(a,d,\alpha,\delta),\psi'(b,c,\beta,\gamma)}$
N	M	$\langle a*_{lo}d, a*_{hi}c\rangle_{\psi'(a,d,\alpha,\delta),\psi'(a,c,\alpha,\gamma)}$
N	N	$\langle b*_{lo}d, a*_{hi}c\rangle_{\psi'(b,d,\beta,\delta),\psi'(a,c,\alpha,\gamma)}$
Z	P, M, N	$\langle 0,0\rangle_{t,t}$
P, M, N, Z	Z	$\langle 0,0\rangle_{t,t}$
any	$\emptyset$	$\emptyset$
$\emptyset$	any	$\emptyset$

FIG. 10. Multiplication of general IEEE intervals. where $\psi'(a, c, \alpha, \gamma) = \eta(a * c) \wedge ((\alpha \wedge \gamma) \vee (\alpha \wedge (a = 0)) \vee (\gamma \wedge (c = 0)))$ and $\eta(x) = \mathit{true}$ if and only if x is a finite IEEE number.

Observe that not every set has an optimal general IEEE approximation, but any set with finitely many connected components will have one. To be able to compute optimal general IEEE approximations, we need to introduce the following boolean function, η.

Definition 16. Let F denote the set of floating point numbers and let $\eta(x)$ be the Boolean function on the extended reals $\mathcal{R}^* = \mathcal{R} \cup \{-\infty, \infty\}$ which is true if $x \in F \cap \mathcal{R}$ and false otherwise. That is, η is the characteristic function of $F \cap \mathcal{R}$.

Observe that $\eta(x * y)$ and $\eta(x/y)$ for floating point numbers x and y can efficiently be computed using the IEEE standard, by checking for the "inexactness exception," which the hardware throws whenever the operations require rounding. This information is precisely what we need to obtain an optimal approximation of the interval arithmetic operations, as the following theorem shows:

THEOREM 16. *Let $X = \langle a, b\rangle_{\alpha,\beta}$ and $Y = \langle c, d\rangle_{\gamma,\delta}$ be general IEEE-standard intervals.*

$$\begin{aligned}\Gamma^*(\langle a, b\rangle_{\alpha,\beta} &+ \langle c, d\rangle_{\gamma,\delta}) \\ &= \langle a +_{lo} c, b +_{hi} d\rangle_{\alpha\wedge\gamma\wedge\eta(a+c),\beta\wedge\delta\wedge\eta(b+d)} \\ \Gamma^*(\langle a, b\rangle_{\alpha,\beta} &- \langle c, d\rangle_{\gamma,\delta}) \\ &= \langle a -_{lo} d, b -_{hi} c\rangle_{\alpha\wedge\delta\wedge\eta(a-d),\beta\wedge\gamma\wedge\eta(b-c)}\end{aligned}$$

and the optimal general IEEE approximations of $X * Y$ and X/Y are given by the formula tables in Figures 10 *and* 11. *These tables make use of the formulas*

$$\begin{aligned}\psi(a, c, \alpha, \gamma) &= (\alpha \wedge \gamma) \vee (\alpha \wedge (a = 0)) \vee (\gamma \wedge (c = 0)) \\ \psi'(a, c, \alpha, \gamma) &= \psi(a, c, \alpha, \gamma) \wedge \eta(a * c) \\ \psi''(a, c, \alpha, \gamma) &= \psi(a, c, \alpha, \gamma) \wedge \eta(a/d)\end{aligned}$$

where $\eta(r)$ is true if and only if r is a finite IEEE number. These formulas also require that $Y = \langle c, d\rangle_{\gamma,\delta}$ adheres to the signed zero convention of Section 5.2, *that is, $c \neq -0$ and $d \neq +0$.*

Class of $\langle a, b\rangle_{\alpha,\beta}$	Class of $\langle c, d\rangle_{\gamma,\delta}$	$\Gamma^*(\langle a, b\rangle_{\alpha,\beta}/\langle c, d\rangle_{\gamma,\delta})$
P	P	$\langle a/_{lo}d, b/_{hi}c\rangle_{\psi''(a,d,\alpha,\delta),\psi''(b,c,\beta,\gamma)}$
M	P	$\langle a/_{lo}c, b/_{hi}c\rangle_{\psi''(a,c,\alpha,\gamma),\psi''(b,c,\beta,\gamma)}$
N	P	$\langle a/_{lo}c, b/_{hi}d\rangle_{\psi''(a,c,\alpha,\gamma),\psi''(b,d,\beta,\delta)}$
P	M	$\langle -\infty, a/_{hi}c\rangle_{f,\psi''(a,c,\alpha,\gamma)} \cup \langle a/_{lo}d, \infty\rangle_{\psi''(a,d,\alpha,\delta),f}$
M	M	$\langle -\infty, +\infty\rangle_{f,f}$
N	M	$\langle -\infty, b/_{hi}d\rangle_{f,\psi''(b,d,\beta,\delta)} \cup \langle b/_{lo}c, \infty\rangle_{\psi''(b,c,\beta,\gamma),f}$
P	N	$\langle b/_{lo}d, a/_{hi}c\rangle_{\psi''(b,d,\beta,\delta),\psi''(a,c,\alpha,\gamma)}$
M	N	$\langle b/_{lo}d, a/_{hi}d\rangle_{\psi''(b,d,\beta,\delta),\psi''(a,d,\alpha,\delta)}$
N	N	$\langle b/_{lo}c, a/_{hi}d\rangle_{\psi''(b,c,\beta,\gamma),\psi''(a,d,\alpha,\delta)}$
Z	P, M, N	$\langle 0, 0\rangle_{t,t}$
any	$Z, \emptyset$	$\emptyset$
$\emptyset$	any	$\emptyset$

FIG. 11. Functional division of general IEEE intervals. Note: This table shows the optimal approximation $\Gamma^*(X/Y)$, that is, the smallest union of general intervals with IEEE endpoints that contains X/Y, where $\psi''(a, c, \alpha, \gamma) = \eta(a/c) \wedge ((\alpha \wedge \gamma) \vee (\alpha \wedge (a=0)) \vee (\gamma \wedge (c=0)))$ and $\eta(x)$ is true if and only if x is a finite IEEE number. These formulas also require that $\langle c, d\rangle_{\gamma,\delta}$ adheres to the signed zero convention of Section 5.2, that is, $c \neq -0$ and $d \neq +0$.

$a=0$	$c=0$	$\psi(a,c,\alpha,\gamma)$	$a*c$	$\eta(a*c)$	$\psi'(a,c,\alpha,\gamma)$	a/c	$\eta(a/c)$	$\psi''(a,c,\alpha,\gamma)$
F	F	$\alpha \wedge \gamma$	$a*c$	$\eta(a*c)$	$\alpha \wedge \gamma \wedge \eta(a*c)$	a/c	$\eta(a/c)$	$\alpha \wedge \gamma \wedge \eta(a/c)$
F	T	γ	0	T	γ	$\pm\infty$	F	F
T	F	α	0	T	α	0	T	α
T	T	$\alpha \vee \gamma$	0	T	$\alpha \vee \gamma$	-	-	-

FIG. 12. Endpoint calculations in multiplication and division of general intervals. Note: These functions are used in Figures 8, 9, 10, and 11 and are never called when the relevant expression $a * c$ or a/c would be NaN, (i.e., $0 * \pm\infty$, $0/0$, ∞/∞, ...).

PROOF. The multiplication formulas in Figure 10 are obtained from the corresponding formulas in Figure 8 for multiplication of general intervals by using outward rounding to guarantee that the resulting interval is an IEEE-standard interval that provides a sound approximation to the result. We then observe that if outward rounding is necessary in computing a given endpoint, then the optimal approximation is obtained by not including that endpoint in the interval. This is indicated by conjoining the Boolean formula ψ for the endpoint $E = a * c$ with $\eta(a * c)$ and results in the formulas in Figure 10. The resulting Boolean formula ψ' is shown in Figure 12.

The IEEE division formulas in Figure 11 arise similarly from the general division formulas in Figure 9 but where we conjoin ψ with $\eta(a/c)$ to get an endpoint formula ψ'' (see Figure 12). However, we can go further; the signed zero convention allows some cases to be combined since it removes the need to check for division by zero. For example, the P/P_1 and P/P_0 rows of Figure 9 are nearly identical in the case of general intervals:

$$X/Y = \begin{cases} \langle a/d, b/c\rangle_{\psi(a,d,\alpha,\delta),\psi(b,c,\beta,\gamma)} & \text{if } c > 0 \\ \langle a/d, \infty\rangle_{\psi(a,d,\alpha,\delta), f} & \text{if } c = 0. \end{cases}$$

In the case of IEEE intervals, if $c = 0$, then we can assume that c is a positive zero by the signed zero convention, and hence $b/c = +\infty$. Thus, this compound rule can be rewritten as the following single rule of Figure 11 for computing the optimal sound IEEE approximation for X/Y:

$$\begin{aligned}\Gamma^*(X/Y) &= \langle a/_{lo}d, b/_{hi}c\rangle_{\psi(a,d,\alpha,\delta)\wedge\eta(a/d),\psi(b,c,\beta,\gamma)\wedge\eta(b/c)}\\ &= \langle a/_{lo}d, b/_{hi}c\rangle_{\psi''(a,d,\alpha,\delta),\psi''(b,c,\beta,\gamma)}\end{aligned}$$

The other cases which can be so compressed are P/N, N/P, N/N, M/P, and M/N and doing so leads to the formulas of Figure 11. □

We end this section with the observation that if U and V are finite unions of general intervals (respectively, general IEEE intervals) U_i and V_j, then for each operator $\circ \in \{+, -, *, /\}$ we can compute $U \circ V$ as the union of the $U_i \circ V_j$, which will again be a finite union of general intervals (respectively, general IEEE intervals) that can be computed from the formulas in this section. Although finite unions of general intervals would most likely be too inefficient to be a good candidate as a practical foundation for interval arithmetic, they could be used as part of an arithmetic expression evaluator. For example, one could allow the values of the subexpressions of a given expression to be represented as finite unions of general intervals, but the final result would be returned as a single (closed) interval by taking the smallest (closed) interval containing the computed result set.

7. *Related Work*

Contributions to division by an interval containing zero span a long period. The initial contribution by Kahan [1968] was important if only for pointing out that such an operation can be usefully defined. Kahan capitalized on the fact that, whether division results in a connected set or not, only a single pair of reals is needed to specify the result. Such a pair is, according to Kahan, an *interior interval* (a connected set) or an *exterior interval* (a union of two connected sets).

Various expressions for the endpoints of the result of interval division of two bounded real intervals can be found in Hammer et al. [1993], Ratz [1996], Novoa [1993], Hansen [1992], and Walster [1998]. The book [Hansen 1992] defines an interval division that returns results between functional and relational division. Except for Ratz [1996], all of these expressions are presented as *defininitions* of division for an extended interval arithmetic and many give inconsistent answers for the case of [0, 0]/[0, 1]. As far as we know, Ratz [1996], was the first to define interval division set theoretically in the case where the denominator contains zero and to prove that his rules are correct, at least for the case of closed and bounded intervals.

The work on BNR Prolog [Older 1989] was in many ways the most advanced version of interval arithmetic when its Unix version came out in 1994. Like Pascal-XSC, BNR Prolog optimally determines that, for example, $\langle -0.5, 0.5\rangle \cap (\langle 1, 1\rangle/\langle -1, 1\rangle)$ is empty. Unlike the Pascal-XSC code in Hammer et al. [1993], BNR Prolog optimally determines $\langle -1, 0\rangle \cap (\langle 1, 1\rangle/\langle 1, \infty\rangle)$ to be empty because the quotient does not contain its greatest lower bound. Unfortunately, apart from Older [1989] and Benhamou and Older [1997], nothing seems to have been published about the arithmetic of BNR Prolog.

8. *Conclusions*

Reals are hard to represent and operate on. Paradoxically, *sets* of reals have tractable and sound approximations. Undefined results of real operations have given trouble in various forms, such as overflow and division by zero. We have shown that these are avoidable.

What most people know about interval arithmetic is that it is a safe alternative for expression evaluation. The practitioner rightly suspects that the usual examples, where expression evaluation goes wildly wrong, are contrived. What is not widely known is that interval arithmetic is a powerful method for extending numerical computation into areas such as nonlinear equation solving and nonconvex global optimization [Hansen 1992; Van Hentenryck et al. 1997], where noninterval methods experience serious difficulties.

We should emphasize that we do not attempt to decree that the full details of interval division, as revealed in this paper, must be implemented in any interval arithmetic system. Indeed, in any implementation, efficiency and simplicity of code have to be weighed against minimizing departures from optimality. Our work can be interpreted as saying to implementers: "Here are *all* the cases in which topologically distinguishable results appear—there *is* no more detail. Preserving correctness, simplify as much as you need to." The formulas for interval arithmetic operations on closed, connected sets of reals have in fact been used successfully in several interval-arithmetic-based constraint solvers [Hickey 2000a, 2000b; Hickey et al. 2000].

In our view, interval arithmetic should be simply and clearly defined in terms of the underlying mathematical model of real arithmetic. Much of scientific and numerical computation is based on a mathematical model in which the variables range over the reals. In conventional computation, floating-point operations are *substituted* for the real arithmetic operations, with all the well-documented dire consequences; see Forsythe [1970] for an early warning.

The current state of the art makes it possible to regard interval computations as computer-generated proofs that certain reals (*real* reals) belong to certain small sets of reals (intervals with floating-point endpoints that are not much wider than the limits imposed by the processor's precision). This breakthrough has taken place by piecemeal improvements through the long history of interval arithmetic.

In this article, we have demonstrated that one can formulate a theory of Interval Arithmetic based on extended notions of intervals that allow intervals to be unbounded and nonclosed, and that this results in an elegant theory that is correct, closed, total, optimal, and efficient.

ACKNOWLEDGMENT. The authors wish to thank Huan Wu for valuable discussions and comments. We are also indebted to the referees whose detailed and insightful comments have greatly contributed to the clarity of the paper.

REFERENCES

ALEFELD, G., AND HERZBERGER, J. 1983. *Introduction to Interval Computations*. Academic Press, Orlando, Fla.

BENHAMOU, F., AND OLDER, W. J. 1997. Applying interval arithmetic to real, integer, and Boolean constraints. *J. Logic Prog.* 32, 1–24.

FORSYTHE, G. E. 1970. Pitfalls of computation, or why a math book isn't enough. *Amer. Math. Monthly 77*, 931–956.

HAMMER, R., HOCKS, M., KULISCH, U., AND RATZ, D. 1993. *Numerical Toolbox for Verified Computing I*. Springer-Verlag, New York.

HANSEN, E. 1992. *Global Optimization Using Interval Analysis*. Marcel Dekker.

HANSON, R. J. 1968. Interval arithmetic as a closed arithmetic system on a computer. Tech. Rep. 197. Jet Propulsion Laboratory.

HICKEY, T. J. 2000a. CLIP: A CLP(Intervals) Dialect for Metalevel Constraint Solving. In *Proceedings of PADL'00*. Lecture Notes in Computer Science, vol. 1753. Springer-Verlag, New York.

HICKEY, T. J. 2000b. Analytic constraint solving and interval arithmetic. In *Proceedings of the 27th Annual ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages* (Jan.). ACM, New York.

HICKEY, T. J., QIU, Z., AND VAN EMDEN, M. H. 2000. Interval constraint plotting for interactive visual exploration of implicitly defined relations. Special issue on Reliable Geometric Computations. *Rel. Comput. 6*, 1.

KAHAN, W. M. 1968. A more complete interval arithmetic. Tech. rep. Univ. Toronto, Toronto, Ont., Canada.

LIPSCHUTZ, S. 1965. *General Topology*. Schaum's Outline Series.

MOORE, R. E. 1966. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, N.J.

NOVOA M. 1993. Theory of preconditioners for the interval Gauss-Seidel method and existence/uniqueness theory with interval Newton methods. Dept. Mathematics, Univ. Southwestern Louisiana.

OLDER, W. J. 1989. Interval arithmetic specification. Tech. rep. Bell-Northern Research Computing Research Laboratory.

RALL, L. B. 1969. Computational solution of nonlinear operator equations. Wiley, New York.

RATZ, D. 1996. On extended interval arithmetic and inclusion isotonicity. Institut für Angewandte Mathematik, Universität Karlsruhe.

STOLFI, J., AND DE FIGUEIREDO, L. 1997. Self-Validated Numerical Methods and Applications. IMPA, Rio de Janiero, Brazil.

VAN HENTENRYCK, P., MICHEL, L., AND DEVILLE, Y. 1997. *Numerica*: *A Modeling Language for Global Optimization*. MIT Press, Cambridge, Mass.

WALSTER, G. W. 1996. The extended real interval system. Available on the internet, http://www.mscs.mu.edu/globsol/readings.html.

WALSTER, G. W., AND HANSEN, E. R. 1998. Interval Algebra, Composite Functions and Dependence in Compilers. Available on the internet, http://www.mscs.mu.edu/globsol/readings.html.

RECEIVED JUNE 1999; REVISED JULY 2001; ACCEPTED JULY 2001