BIT Numerical Mathematics, 52(4):1059–1075, 2012

INTERVAL ARITHMETIC OVER FINITELY MANY ENDPOINTS

SIEGFRIED M. RUMP *

Abstract. To my knowledge all definitions of interval arithmetic start with real endpoints and prove properties. Then, for practical use, the definition is specialized to finitely many endpoints, where many of the mathematical properties are no longer valid. There seems no treatment how to choose this finite set of endpoints to preserve as many mathematical properties as possible.

Here we define interval endpoints directly using a finite set which, for example, may be based on the IEEE 754 floating-point standard. The corresponding interval operations emerge naturally from the corresponding power set operations. We present necessary and sufficient conditions on this finite set to ensure desirable mathematical properties, many of which are not satisfied by other definitions. For example, an interval product contains zero if and only if one of the factors does.

The key feature of the theoretical foundation is that "endpoints" of intervals are not points but non-overlapping closed, half-open or open intervals, each of which can be regarded as an atomic object. By using non-closed intervals among its "endpoints", intervals containing "arbitrarily large" and "arbitrarily close to but not equal to" a real number can be handled. The latter may be zero defining "tiny" numbers, but also any other quantity including transcendental numbers.

Our scheme can be implemented straightforwardly using the IEEE 754 floating-point standard.

Key words. Interval arithmetic IEEE 754 finitely many endpoints mathematical properties

AMS subject classifications. 65G30

1. Introduction. We assume that the reader is familiar with basic concepts of interval arithmetic as can be found, for example, in [5, 7, 11]. The aim of this paper is to define an interval arithmetic over a finite set of possible endpoints preserving as much mathematical properties as possible. There are natural limitations due to the finiteness of the set of possible bounds which, in general, do not permit associativity of interval addition and multiplication (see Section 3). Nevertheless we can preserve properties which are not satisfied by any other definition including interval arithmetics with infinitely many endpoints.

Usually (cf. [5, 7, 11]) intervals $[\alpha,\beta]$ are defined in a first step with endpoints $\alpha,\beta\in\mathbb{R}$, and interval operations are defined to be the narrowest (interval) inclusion of the corresponding power set operation. This is often called the *inclusion principle* [11]. To be applicable on digital computers, only endpoints out of a finite set, e.g. of floating-point numbers can be allowed. The definitions are then adapted such that the inclusion principle is not sacrificed.

Unbounded intervals and thus infinite bounds are mandatory [10] to ensure that interval operations are closed. However, it is preferable to define intervals to be always subsets of $\mathbb{R}$, so that unbounded intervals are restricted to half-open intervals $(-\infty,\alpha]$ and $[\alpha,\infty)$, and to $(-\infty,\infty)$. Hence the appealing mathematical property $0\cdot A=0$ is true for all intervals A. This principle is pursued in the proposal of an IEEE interval arithmetic standard [6].

For an arbitrarily large number there is always an interval containing it, and an arbitrarily small positive number is bounded below by zero. In normal interval arithmetic this implies the drawback that $B:=1/A$ is well-defined for $A:=[1,\infty)$, but the left bound of B is zero, so that $1/(1/A)$ contains $1/0$. It seems natural to allow for intervals being half-open at zero, so that, for example, $1/A=(0,1]$ and $1/(1/A)=[1,\infty)=A$.

Although zero is the commonest case, a similar problem occurs for the inverse hyperbolic tangent of $\tanh[0,\infty)=[0,1)$ at $\xi=1$, or for the tangent of $\arctan[0,\infty)=[0,\pi/2)$ at $\xi=\pi/2$.

*Institute for Reliable Computing, Hamburg University of Technology, Schwarzenbergstraße 95, 21071 Hamburg, Germany and Visiting Professor at Waseda University, Faculty of Science and Engineering, 3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, Japan, rump@tu-harburg.de

Therefore we aim to address this problem by allowing intervals being half-open at a general, even transcendental point $\xi \in \mathbb{R}$.

In the following we do not restrict ourselves to such specific examples, but present a general scheme to define an interval arithmetic over a finite set of interval bounds. Necessary and sufficient conditions are developed under which desirable mathematical properties are true, for example $A \subseteq 1/(1/A)$ for any interval A not containing zero.

Rather than defining intervals first using the infinite set of real endpoints, we define intervals directly over a finite set of endpoints. Exactly speaking our intervals are the convex union of two intervals, so that the endpoints of our intervals are formally defined to be intervals themselves. As we will see this formalism resolves the mentioned shortcomings of normal interval arithmetic.

One particular choice of interval endpoints is based on floating-point numbers, for instance according to the IEEE 754 floating-point standard (cf. [1, 2]). Although most of the interval computations can be performed using floating-point operations with directed rounding, some care is necessary concerning infinite bounds. For example, IEEE 754 defines $0 \cdot \infty = \texttt{NaN}$, so that multiplication of an interval bound by zero results not necessarily in zero. Note that there are other definitions; for example, sometimes (but not always) $0 \cdot \infty = 0$ is used in measure theory [3].

A problem of conventional definitions [6] is that an infinite bound is not a member of its own interval. In particular, the intersection of the interval $A := [+\infty, +\infty]$ with $\mathbb{R}$ is empty. By defining A to be empty, seemingly natural properties such as

$$[\alpha, \beta] = \text{hull}(\text{interval}(\alpha), \text{interval}(\beta)) \qquad \text{or} \qquad \alpha \in \text{interval}(\alpha) \tag{1.1}$$

are not valid any more. There are circumventions in [6] by defining other access functions to interval bounds. Nevertheless, I think it is hazardous to define an interval arithmetic with bounds out of a set $\mathbb{B}$ for which there are elements in $\mathbb{B}$ not belonging to any interval.

A solution is to interpret an infinite bound as "huge", not infinity but in a way larger than any real number. Examples for a rigorous treatment of such ideas are surreal numbers [4] or nonstandard analysis [9]. This, however, is far too much for our purposes.

To start with a finite set $\mathbb{B}$ of endpoints and then to define an interval arithmetic directly over $\mathbb{B}$ rather than first over $\mathbb{R}$ avoids not only the mentioned problems but makes it also directly suitable for a computer implementation. We give necessary and sufficient conditions on $\mathbb{B}$ so that the new interval arithmetic preserves mathematical properties such as

$$\begin{aligned} &0 \in A - B \;\Leftrightarrow\; A \cap B \neq \emptyset \\ &0 \in A \cdot B \;\Leftrightarrow\; 0 \in A \cup B \\ &A \subseteq B/(B/A) \quad \text{if } 0 \notin A \cup B \end{aligned}$$

avoiding problems with underflow, and

$$\begin{aligned} &0 \cdot A = [0, 0] \\ &\alpha \in \text{interval}(\alpha) \\ &[\alpha, \beta] = \text{hull}(\text{interval}(\alpha), \text{interval}(\beta)) \end{aligned}$$

avoiding problems with overflow and infinity mentioned before. Other desirable properties such as

$$A \subseteq \log(\exp(A)) \quad \text{for any interval } A \tag{1.2}$$

are valid as well. Basically, it is necessary and sufficient to introduce quantities "huge" and "tiny" to be endpoints, but avoiding infinity to be an endpoint or element of an interval.

We do not know of another interval arithmetic with the above properties. Consider, for example, the set of intervals $\{x \in \mathbb{R} : a \le x \le b\}$ for $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$ as in [6]. Note that the bounds are *real* numbers. If an interval A is unbounded to the left, then in normal interval arithmetic the best lower bound for $B := \exp(A)$ is zero, so that $\log(B)$ is not defined for all $b \in B$. Hence even for intervals with *real* endpoints, (1.2) is not satisfied for the standard definitions [5, 7, 6, 11] of interval arithmetic.[1]

Rather than defining a specific interval arithmetic, we develop and investigate a general scheme and analyze it. This covers in particular the above mentioned, but allows much more.

2. Notation and definitions. We start with a theoretical treatment which will turn out to be directly suitable for a practical implementation.

The purpose of this paper is to define intervals over finitely many endpoints, such as floating-point numbers, and an interval arithmetic. To avoid confusion with ordinary intervals such as $[\alpha, \beta]$, $[\alpha, \beta)$ etc. over real numbers, we call the latter $\mathbb{R}$-intervals. Thus, the set $\mathbb{IR}$ of $\mathbb{R}$-intervals is the set of non-empty and connected subsets of $\mathbb{R}$. This covers in particular unbounded $\mathbb{R}$-intervals.

The following definition identifies $\mathbb{R}$-intervals to be the *bounds* of our to-be-defined new intervals. Seemingly strange at first sight, this formalizes what we want to do. In a practical implementation those $\mathbb{R}$-intervals would basically consist of the set of degenerated intervals $[f, f]$ for all floating-point numbers f plus some extra bounds to be specified.

DEFINITION 2.1. *A finite set* $\mathbb{B} = \{b_1, \ldots, b_k\}$ *is called a* weakly admissible set of interval bounds *if* $b_i \in \mathbb{IR}$ *for all* $1 \le i \le k$ *and*

$$\alpha \in b_i, \beta \in b_{i+1} \quad \Rightarrow \quad \alpha < \beta \qquad \text{for } 1 \le i < k\,. \tag{2.1}$$

If, in addition, $k > 1$ *and*

$$\inf b_1 = -\infty \qquad \text{and} \qquad \sup b_k = +\infty\,, \tag{2.2}$$

then $\mathbb{B}$ *is called an* admissible set of interval bounds.

REMARK 2.2. *The condition* $k > 1$ *excludes the trivial case* $\mathbb{B} = \{b_1\}$ *with* $b_1 = \mathbb{R}$ *to be an admissible set of interval bounds.*

EXAMPLE 2.3. *Let* $F = \{f_1, \ldots, f_n\} \subset \mathbb{R}$ *be a finite set of real numbers, let* $\mu \in \mathbb{R}$ *be such that* $|f| \le \mu$ *for all* $f \in F$*, and define* $H := \{\alpha \in \mathbb{R} : \mu < \alpha\}$*. Then* $\mathbb{B} := \{\{f\} : f \in F\} \cup \{-H, H\}$ *is an admissible set of interval bounds. Note that this is also true for replacing* H *by* $\{\alpha \in \mathbb{R} : \mu + 1 \le \alpha\}$.

An individual element $b \in \mathbb{B}$ may be a set consisting of a single real number, or may be an open, a half-open or a closed $\mathbb{R}$-interval. In particular, b_1 and b_k may be unbounded. Note that the $b_i \in \mathbb{B}$ are mutually disjoint $\mathbb{R}$-intervals, but will serve as bounds for our intervals to be defined.

We define a total ordering $\preceq$ on $\mathbb{B}$ by

$$b_i \preceq b_j \quad \text{for } 1 \le i \le j \le k\,. \tag{2.3}$$

Furthermore $a \prec b$ means $a \preceq b$ and $a \ne b$,

$$\min\nolimits_{\preceq}(a, b) := \begin{cases} a & \text{if } a \preceq b \\ b & \text{otherwise} \end{cases} \tag{2.4}$$

[1] Another definition of interval arithmetic by Hansen and Walster with theoretical foundation by Pryce, cf. [8], uses *containment sets* (cset), basically "ignoring input out of range", see Section 5. In that case (1.2) is satisfied. However, $B \not\subseteq \exp(\log(B))$ for $B = [-1, 1]$ because $\log B = (-\infty, 0]$ in that arithmetic.

for $a, b \in \mathbb{B}$, and similarly $\max_{\preceq}(a,b)$.

DEFINITION 2.4. *The set* $\mathbb{IB}$ *of* proper intervals *over a weakly admissible set of interval bounds* $\mathbb{B}$ *is the empty set and the set of all pairs* $(a,b) \in \mathbb{B} \times \mathbb{B}$ *with* $a \preceq b$. To avoid confusion with $\mathbb{R}$-intervals we use the notation $[\![a,b]\!]$, so that

$$\mathbb{IB} = \{[\![a,b]\!] : a,b \in \mathbb{B},\ a \preceq b\} \cup \{\emptyset\} . \tag{2.5}$$

Note that the notation implies that a proper interval $[\![a,b]\!]$ is non-empty, and that $a \preceq b$.

DEFINITION 2.5. *We call the convex union* $a \underline{\cup} b$ *of* a *and* b *the range of the proper interval* $[\![a,b]\!] \in \mathbb{IB}$*, i.e. range*$([\![a,b]\!]) := a \underline{\cup} b \subseteq \mathbb{R}$. *Moreover, range*$(\emptyset) := \emptyset$. *We use an auxiliary quantity* `NaI` *(Not an Interval) to define the set* $\overline{\mathbb{IB}}$ *of* intervals *by*

$$\overline{\mathbb{IB}} = \mathbb{IB} \cup \{\mathtt{NaI}\} . \tag{2.6}$$

For $\mathbb{B} = \{b_1, \ldots, b_k\}$*, the quantity*

$$\underline{\bigcup} \{b : b \in \mathbb{B}\} = b_1 \underline{\cup} b_k \subseteq \mathbb{R} \tag{2.7}$$

is called the range *of intervals.*

The range of intervals is $\mathbb{R}$ if and only if $\mathbb{B}$ is admissible. The range of `NaI` is not defined. The quantity `NaI` will serve as the default result of an interval operation if one operand is `NaI`, or if some input is out of range (such as division by zero). Thus we concentrate in the following on the definition of interval operations on proper interval. There are other ways to handle such exceptions, see Section 5.

Set operations on proper intervals are defined by identifying the interval with its range, for example

$$\xi \in [\![a,b]\!] \ :\Leftrightarrow\ \xi \in a \underline{\cup} b \qquad \text{for } \xi \in \mathbb{R} \tag{2.8}$$

and

$$a = \mathrm{range}([\![a,a]\!]) \qquad \text{for } a \in \mathbb{B} . \tag{2.9}$$

Moreover,

$$[\![a,b]\!] \subseteq [\![c,d]\!] \ :\Leftrightarrow\ a \underline{\cup} b \subseteq c \underline{\cup} d \qquad \text{for } [\![a,b]\!], [\![c,d]\!] \in \mathbb{IB} . \tag{2.10}$$

Note that, although the endpoints are out of a finite set $\mathbb{B}$, a proper interval covers all *real* numbers in its range.

If the intersection of two proper intervals $[\![a,b]\!]$ and $[\![c,d]\!]$ is not empty, then

$$[\![a,b]\!] \cap [\![c,d]\!] = [\![\max_{\preceq}(a,c), \min_{\preceq}(b,d)]\!] , \tag{2.11}$$

whence $\mathbb{IB}$ is closed under intersection. The hull always satisfies

$$\mathrm{hull}([\![a,b]\!], [\![c,d]\!]) = [\![\min_{\preceq}(a,c), \max_{\preceq}(b,d)]\!] . \tag{2.12}$$

In particular for all $a, b \in \mathbb{B}$,

$$[\![\min_{\preceq}(a,b), \max_{\preceq}(a,b)]\!] = \mathrm{hull}([\![a,a]\!], [\![b,b]\!]) \quad \text{with range } a \underline{\cup} b . \tag{2.13}$$

COROLLARY 2.6. *The set* $\mathbb{IB}$ *of proper intervals over a weakly admissible set of interval bounds* $\mathbb{B}$ *forms a complete lattice.*

DEFINITION 2.7. *Interval operations* $\circ : \mathbb{IB} \times \mathbb{IB} \to \mathbb{IB}$ *for* $\circ \in \{+,-,\cdot,/\}$ *on proper intervals are defined by*

$$A \circ B := \bigcap \{C \in \mathbb{IB} : \alpha \circ \beta \in C \text{ for all } \alpha \in A, \beta \in B\} \in \mathbb{IB} \tag{2.14}$$

provided $0 \notin C$ *in case of division. This extends to operations* $\circ : \overline{\mathbb{IB}} \times \overline{\mathbb{IB}} \to \overline{\mathbb{IB}}$ *on intervals by*

$$A \circ B := \texttt{NaI} \qquad \text{if } A = \texttt{NaI} \text{ or } B = \texttt{NaI} \text{ or } 0 \in B \text{ in case of division .} \tag{2.15}$$

The intersection in (2.14) is taken over a finite set of proper intervals. It is thus well-defined and again a proper interval. Note that α, β in (2.14) run over all *real* $\alpha \in A$ and $\beta \in B$, so that $\alpha \circ \beta \in \mathbb{R}$ is the real operation between α and β. The result of an interval operation may be empty for weakly admissible $\mathbb{B}$, but, as we will see, not for admissible $\mathbb{B}$.

EXAMPLE 2.8. *Define* $\mathbb{F} := \{m/1000 : m \in \mathbb{Z}, -10^{12} < m < 10^{12}\}$. *Then* $\mathbb{B} := \{\{f\} : f \in \mathbb{F}\}$ *is a weakly admissible set of interval bounds. Identifying* $f \in \mathbb{F}$ *with* $\{f\}$ *one has* $[\![2,2]\!]/[\![3,3]\!] = [\![0.666, 0.667]\!]$, *but* $[\![10^6, 10^6]\!] \cdot [\![10^6, 10^6]\!] = \emptyset$.

In order to allow interval operations with real numbers we define the mapping $\diamondsuit : \mathbb{R} \to \mathbb{IB}$ by

$$\diamondsuit(\xi) := \bigcap \{C \in \mathbb{IB} : \xi \in C\} . \tag{2.16}$$

Note that $\diamondsuit$ is defined for all $\xi \in \mathbb{R}$, and $x \in \diamondsuit(x)$ if and only if $\xi \in b_1 \cup b_k$. If ξ is not in the range of intervals, then $\diamondsuit(\xi) = \emptyset$. Using this embedding, operations $\circ \in \{+,-,\cdot,/\}$ between a real number ξ and an interval A are defined by

$$\xi \circ A := \diamondsuit(\xi) \circ A \qquad \text{and} \qquad A \circ \xi := A \circ \diamondsuit(\xi) . \tag{2.17}$$

One of the most important properties of any interval arithmetic is the inclusion principle, i.e. to cover the range of the power set operations. This applies not only to operations on intervals but also between real numbers. Thus the inclusion monotonicity $\alpha \circ \beta \in \diamondsuit(\alpha) \circ \diamondsuit(\beta)$ for all $\alpha, \beta \in \mathbb{R}$ is most desirable if not mandatory.

THEOREM 2.9. *Let* $\mathbb{B}$ *be a weakly admissible set of interval bounds. Then*

$$\alpha \circ \beta \in \diamondsuit(\alpha) \circ \diamondsuit(\beta) \qquad \text{for } \circ \in \{+,-,\cdot\} \text{ and all } \alpha, \beta \in \mathbb{R} \tag{2.18}$$

is true if and only if $\mathbb{B}$ *is an admissible set of interval bounds, i.e. the range of intervals is* $\mathbb{R}$.

REMARK 2.10. *Note that the assertion is not true for division because* $\diamondsuit(\beta)$ *may contain zero for nonzero* β.

Proof. If the range of intervals is $\mathbb{R}$, then $\xi \in \diamondsuit(x)$ for all $\xi \in \mathbb{R}$. Moreover, $\alpha \circ \beta$ is defined for any real α, β, so that it cannot happen that $\alpha \circ \beta$ is defined but $\alpha' \circ \beta'$ is not for $\alpha' \in \diamondsuit(\alpha)$ and $\beta' \in \diamondsuit(\beta)$. Hence (2.18) follows by the definition (2.14).

If, for $\mathbb{B} = \{b_1, \ldots, b_k\}$, the supremum of b_k is $\sigma < \infty$, then $\diamondsuit(\xi) = \emptyset$ for $\xi := \sigma + 1$ and therefore $\diamondsuit(\xi) + \diamondsuit(\xi) = \emptyset$, a contradiction to (2.18). If the infimum of b_1 is finite, we proceed similarly. ∎

Because of the utmost importance of the inclusion monotonicity (2.18), we focus our attention in the following on admissible sets of interval bounds. In this case the range of intervals is $\mathbb{R}$, and the smallest element b_1 and largest element b_k in the ordering $\preceq$ may be interpreted as a representation of the overflow range.

DEFINITION 2.11. *For a function* $f : D_f \subseteq \mathbb{R}^n \to \mathbb{R}$ *its* natural interval extension $F : \overline{\mathbb{IB}}^n \to \overline{\mathbb{IB}}$ *is defined by*

$$F(A) := \begin{cases} \bigcap \{C \in \mathbb{IB} : f(\alpha) \in C \text{ for all } \alpha \in A\} & \text{if } A \in \mathbb{IB}^n \text{ and } A \subseteq D_f \\ \texttt{NaI} & \text{otherwise ,} \end{cases} \tag{2.19}$$

where for $\alpha := (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ *and* $A := (A_1, \ldots, A_n) \in \mathbb{IB}^n$ *we use the vector notation*

$$\alpha \in A \;:\Leftrightarrow\; \alpha_i \in A_i \text{ for all } 1 \le i \le n \quad \text{and} \quad A \subseteq D_f :\Leftrightarrow \alpha \in D_f \text{ for all } \alpha \in A\,. \tag{2.20}$$

A function $G : \overline{\mathbb{IB}}^n \to \overline{\mathbb{IB}}$ *is called a* weak interval extension *of* f *if, for all* $A \in \overline{\mathbb{IB}}^n$,

$$G(A) \neq \mathtt{NaI} \;\Rightarrow\; F(A) \subseteq G(A)\,. \tag{2.21}$$

REMARK 2.12. *Definition 2.7 implies that* $F(A,B) := A \circ B$ *is the natural interval extension of* $f(\alpha,\beta) := \alpha \circ \beta$ *for* $\circ \in \{+,-,\cdot,/\}$.

We close this section with two simple examples. The set $\mathbb{B} := \{N, P_0\}$ with

$$N := \{\xi \in \mathbb{R} : \xi < 0\} \qquad \text{and} \qquad P_0 := \{\xi \in \mathbb{R} : \xi \ge 0\} \tag{2.22}$$

is an admissible set of interval bounds. For positive $\pi \in \mathbb{R}$ one has $\diamondsuit(\pi) = [\![P_0, P_0]\!]$, and therefore

$$\diamondsuit(\xi)/\diamondsuit(\pi) = \mathtt{NaI} \qquad \text{for any } \; \xi \in \mathbb{R} \text{ and } 0 < \pi \in \mathbb{R}\,. \tag{2.23}$$

The set $\mathbb{B} := \{N, P\}$ with

$$P := \{\xi \in \mathbb{R} : \xi > 0\} \tag{2.24}$$

is also an admissible set of interval bounds. One has $\diamondsuit(\pi) = [\![P,P]\!]$ for positive $\pi \in \mathbb{R}$, and $\diamondsuit(\eta) = [\![N,N]\!]$ for negative $\eta \in \mathbb{R}$, whereas $\diamondsuit(0) = [\![N,P]\!]$. Thus

$$\diamondsuit(\alpha)/\diamondsuit(\beta) = \begin{cases} [\![P,P]\!] & \text{if } \; \alpha\beta > 0 \\ [\![N,N]\!] & \text{if } \; \alpha\beta < 0 \\ [\![N,P]\!] & \text{if } \; \alpha = 0,\, \beta \neq 0 \\ \mathtt{NaI} & \text{if } \; \beta = 0\,. \end{cases} \tag{2.25}$$

Hence $\alpha/\beta \in \diamondsuit(\alpha)/\diamondsuit(\beta)$ for all $\alpha, \beta \in \mathbb{R}$ with $\beta \neq 0$, thus extending Theorem 2.9 to division. In the next section we will characterize under which circumstances this and other desirable properties are true.

3. Mathematical properties. As has been mentioned in the introduction, the finiteness of the set of bounds $\mathbb{B}$ does not permit associativity of interval addition and multiplication under general assumptions.

THEOREM 3.1. *Let an admissible set of interval bounds* $\mathbb{B}$ *be given. If* $\{0\}, \{\alpha\} \in \mathbb{B}$ *for* $0 < \alpha \in \mathbb{R}$*, then interval addition is not associative. If* $\{1/\alpha\}, \{1\}, \{\alpha\} \in \mathbb{B}$ *for* $1 < \alpha \in \mathbb{R}$*, then interval multiplication is not associative.*

Proof. Since $\mathbb{B}$ is admissible, b_1 and b_k are not bounded. Define $A := [\![\{\alpha\}, \{\alpha\}]\!]$ and $B := [\![b_{k-1}, b_{k-1}]\!]$. Then $\{\alpha\} \preceq b_{k-1}$ and $B + (A - A) = B \neq (B + A) - A$.

If $\{1/\alpha\}, \{1\}, \{\alpha\} \in \mathbb{B}$, then again $\{\alpha\} \preceq b_{k-1}$. Define A and B as before, and $C := [\![\{\alpha\}, \{\alpha\}]\!]$ and $D := [\![\{1/\alpha\}, \{1/\alpha\}]\!]$. Then $B(CD) = B \neq (BC)D$. ∎

Both examples use unbounded intervals as intermediate results. At least for addition they have to: For $\mathbb{B}$ based on a fixed-point number system such as $\{k\alpha : k \in \mathbb{Z}, |k| \le K\}$ for nonzero $\alpha \in \mathbb{R}$ appended with bounds covering overflow, addition is exact (and therefore associative) provided no "overflow" occurs.

In Section 4 we show that for $\mathbb{B}$ based on IEEE 754 floating-point numbers, neither addition nor multiplication is associative even if all quantities and intermediate results are finite, and the sub-distributivity is not satisfied as well.

A number of mathematical properties are true. Some of the following properties are called "trivial implications" in [7], page 21. This is indeed true for intervals with *real* endpoints. Here, however, we are restricted to a finite set of bounds, such as floating-point numbers. We do not know of another definition of interval arithmetic maintaining the following properties.

We defined the range of intervals to be a subset of $\mathbb{R}$, not allowing infinity as an element of an interval. Thus multiplication by zero results is zero:

THEOREM 3.2. *Let $A, B \in \mathbb{IB}$ for a weakly admissible set of interval bounds $\mathbb{B}$, and assume $\{0\} \in \mathbb{B}$. Then, identifying 0 with $\{0\}$,*

$$A \cdot B = [\![0,0]\!] \qquad \Leftrightarrow \qquad A = [\![0,0]\!] \quad or \quad B = [\![0,0]\!] \; . \tag{3.1}$$

Proof. Follows directly by the definition (2.14). □

One of the most important properties of any interval arithmetic is the inclusion monotonicity:

THEOREM 3.3. *Let $A, B \in \mathbb{IB}$ for an admissible set of interval bounds $\mathbb{B}$. Then*

$$\alpha \circ \beta \in A \circ B \qquad for\ all\ \ \alpha \in A, \beta \in B \tag{3.2}$$

for $\circ \in \{+, -, \cdot, /\}$, where $0 \notin B$ is assumed in case of division. For $A', B' \in \mathbb{IB}$ we have

$$A \subseteq A',\ B \subseteq B' \qquad \Rightarrow \qquad A \circ B \subseteq A' \circ B' \; , \tag{3.3}$$

where $0 \notin B'$ is assumed in case of division. The assertions (3.2) and (3.3) are not necessarily true for a weakly admissible set of interval bounds $\mathbb{B}$.

For a function $f : D_f \subseteq \mathbb{R}^n \to \mathbb{R}$ and its natural interval extension $F : \overline{\mathbb{IB}}^n \to \overline{\mathbb{IB}}$ one has

$$f(\alpha) \in F(A) \qquad for\ all\ \ \alpha \in A \tag{3.4}$$

provided $A \subseteq D_f$. For a weak interval extension $G : \overline{\mathbb{IB}}^n \to \overline{\mathbb{IB}}$ of f we have

$$G(A) \neq \mathtt{NaI} \quad \Rightarrow \quad f(\alpha) \in G(A)\ for\ all\ \ \alpha \in A \; . \tag{3.5}$$

Proof. Since the range of intervals is $\mathbb{R}$, the assertions follow directly by the definitions (2.14) and (2.19). For a weakly admissible set of interval bounds $\mathbb{B}$ the range of intervals is not necessarily $\mathbb{R}$, so that $A \circ B$ and $A' \circ B'$ may be empty although the power set operation is well-defined. □

We note that Theorem 3.3 extends in an obvious way to arbitrary expressions involving arithmetic operations and functions, in one or in several dimensions. In general, this results in a weak interval extension.

DEFINITION 3.4. *A weakly admissible set of interval bounds $\mathbb{B}$ is called* dense around *$\rho \in \mathbb{R}$ if there are $t_1, t_2 \in \mathbb{B}$, $t_1 \neq t_2$ with*

$$\sup t_1 = \inf t_2 = \rho \qquad and \qquad \rho \notin t_1 \cup t_2 \; . \tag{3.6}$$

REMARK 3.5. *The definition implies $t_1 \preceq t_2$, and $\{\rho\}$ may be an element of $\mathbb{B}$ or not.*

With this definition we can characterize the conditions on $\mathbb{B}$ so that the inclusion monotonicity (2.18) in Theorem 2.9 is also true for division.

THEOREM 3.6. *Let* $\mathbb{B}$ *be an admissible set of interval bounds, i.e. the range of intervals is* $\mathbb{R}$*. Let* $\rho \in \mathbb{R}$ *be given. Then*

$$\rho \neq \xi \quad \Leftrightarrow \quad \rho \notin \diamondsuit(\xi) \qquad \text{for all } \xi \in \mathbb{R} \tag{3.7}$$

is true if and only if $\mathbb{B}$ *is dense around* ρ*. Furthermore,*

$$\alpha \circ \beta \in \diamondsuit(\alpha) \circ \diamondsuit(\beta) \quad \begin{array}{l} \text{for } \circ \in \{+,-,\cdot,/\} \text{ and all } \alpha,\beta \in \mathbb{R}, \\ \beta \neq 0 \text{ in case of division}, \end{array} \tag{3.8}$$

is true if and only if $\mathbb{B}$ *is dense around* 0*.*

Proof. Since $\mathbb{B}$ is admissible, $\xi \in \diamondsuit(\xi)$ for all $\xi \in \mathbb{R}$ so that the direction "$\Leftarrow$" in (3.7) is always true.

If $\mathbb{B}$ is dense around ρ, then (3.7) follows. Suppose $\mathbb{B}$ is not dense around ρ and let $[\![a,b]\!] := \diamondsuit(\rho)$. If the range of $[\![a,b]\!]$ is $\{\rho\}$, then $\sup t_1 \neq \rho$ for all t_1 with $t_1 \prec a$ or $\inf t_2 \neq \rho$ for all t_2 with $b \prec t_2$, so that in either case there is $\beta \neq \rho$ with $\rho \in \diamondsuit(\xi)$. Otherwise, the interior of $[\![a,b]\!]$ is not empty and again there is $\xi \neq \rho$ with $\rho \in \diamondsuit(\xi)$. This proves the first part, and the second part follows by Theorem 2.9 and the equivalence (3.7). ∎

THEOREM 3.7. *Let* $A,B \in \mathbb{IB}$ *for an admissible set of interval bounds* $\mathbb{B}$*. Then for* $A,B \neq \emptyset$ *the equivalence*

$$0 \in A \cdot B \qquad \Leftrightarrow \qquad 0 \in A \quad \textit{or} \quad 0 \in B\,. \tag{3.9}$$

is true if $\mathbb{B}$ *is dense around* 0*.*

REMARK 3.8. *The equivalence may also be true if* $\mathbb{B}$ *is not dense around* 0 *such as for* $\mathbb{B} := \{b_1,b_2\}$ *with* $b_1 := \{\xi \in \mathbb{R} : \xi \leq -1\}$ *and* $b_2 := \{\xi \in \mathbb{R} : \xi \geq 1\}$*. Then* $\mathbb{IB}$ *consists only of the three intervals* $N := [\![b_1,b_1]\!]$*,* $R := [\![b_1,b_2]\!]$ *and* $P := [\![b_2,b_2]\!]$ *with* $N \cdot N = P \cdot P = P$ *and* $R \cdot N = R \cdot P = R$*.*

Proof. The direction "$\Leftarrow$" is trivial. Assume $\mathbb{B} = \{b_1,\ldots,b_k\}$ is dense around 0 with t_1 and t_2 as in Definition 3.4. If $0 \notin A$ and $0 \notin B$, then $M := \{\alpha\beta : \alpha \in A, \beta \in B\}$ satisfies

$$\text{either} \quad M \subseteq \{\xi \in \mathbb{R} : \xi < 0\} \subseteq b_1 \cup t_1 \quad \text{or} \quad M \subseteq \{\xi \in \mathbb{R} : \xi > 0\} \subseteq t_2 \cup b_k\,. \tag{3.10}$$

In either case the definition (2.14) implies $0 \notin A \cdot B$, and therefore (3.9). ∎

THEOREM 3.9. *Let* $A,B \in \mathbb{IB}$ *for an admissible set of interval bounds* $\mathbb{B}$*. If* $\{\xi \in \mathbb{R} : \xi \leq 0\} \notin \mathbb{B}$*,* $B \neq \emptyset$ *and* $0 \notin B$*, then the equivalence*

$$0 \in A/B \qquad \Leftrightarrow \qquad 0 \in A \tag{3.11}$$

is true if and only if $\mathbb{B}$ *is dense around* 0*.*

REMARK 3.10. *The assumption* $N_0 := \{\xi \in \mathbb{R} : \xi \leq 0\} \notin \mathbb{B}$ *excludes the pathological case that* $\diamondsuit(\xi) = [\![N_0,N_0]\!]$ *for each non-positive* ξ*. For example,* $\mathbb{B} := \{N_0,P\}$ *with* $P := \{\xi \in \mathbb{R} : \xi > 0\}$ *is an admissible set of interval bounds which is not dense around* 0*, but for which the equivalence (3.11) holds true.*

Proof. Again, the direction "$\Leftarrow$" in (3.11) is always true, and if $\mathbb{B} = \{b_1,\ldots,b_k\}$ is dense around 0 then the direction "$\Rightarrow$" follows as in the proof of Theorem 3.7. Assume $\mathbb{B}$ is not dense around 0. Then $k > 1$ implies that

$$\exists \alpha_1 \forall \xi \in b_1 :\ \xi \leq \alpha_1 < 0 \qquad \text{or} \qquad \exists \alpha_2 \forall \xi \in b_k :\ 0 < \alpha_2 \leq \xi\,, \tag{3.12}$$

and by assumption both b_1 and b_k are unbounded. Define $A := [\![b_1,b_1]\!]$ and $B := [\![b_k,b_k]\!]$. In the first case, $0 \notin A$ and

$$\{\alpha/\beta : \alpha,\beta \in A\} = P := \{\xi \in \mathbb{R} : \xi > 0\} \subseteq A/A\,, \tag{3.13}$$

and similarly $0 \notin B$ and

$$\{\alpha/\beta : \alpha,\beta \in B\} = P \subseteq B/B \tag{3.14}$$

in the second case. If there is no $b \in \mathbb{B}$ with $\inf b = 0$ and $0 \notin b$, then $0 \in A/A$ or $0 \in B/B$, respectively, a contradiction to (3.11).

On the contrary, assume there exists $b \in \mathbb{B}$ with $\inf b = 0$ and $0 \notin b$. Then $b \preceq b_k$ implies $\xi > 0$ for all $\xi \in B$. Since $b_1 \preceq b$ and the range of intervals is $\mathbb{R}$, we have $\sup b_1 \leq 0$, and $\{\xi \in \mathbb{R} : \xi \leq 0\} \notin \mathbb{B}$ implies that $\alpha < 0$ for all $\alpha \in A$. Hence $\{\alpha/\beta : \alpha \in A, \beta \in B\} = \{\xi \in \mathbb{R} : \xi < 0\}$. But, because $\mathbb{B}$ is not dense around 0, there is no $c \in \mathbb{B}$ with $\sup c = 0$ and $0 \notin c$, and this means $0 \in A/B$. □

THEOREM 3.11. *Let $A, B \in \mathbb{IB}$ for an admissible set of interval bounds $\mathbb{B}$. If $\mathbb{B}$ is dense around 0, then*

$$0 \in A - B \qquad \Leftrightarrow \qquad A \cap B \neq \emptyset\,. \tag{3.15}$$

REMARK 3.12. *The equivalence may also be true if $\mathbb{B}$ is not dense around 0 such as for $\mathbb{B} := \{b_1, b_2\}$ as in Remark 3.8. Then N and P are the only intervals with empty intersection, and neither $N - P = N$ nor $P - N = P$ contain zero.*

Proof. If $\xi \in A \cap B$, then $0 = \xi - \xi \in A - B$, so the direction "$\Leftarrow$" is always true. Suppose $A \cap B = \emptyset$, and assume $\mathbb{B} = \{b_1, \ldots, b_k\}$ is dense around 0 with t_1 and t_2 as in Definition 3.4. Then $M := \{\alpha - \beta : \alpha \in A, \beta \in B\}$ satisfies

$$\text{either} \quad M \subseteq \{\xi \in \mathbb{R} : \xi < 0\} \subseteq b_1 \supseteq t_1 \quad \text{or} \quad M \subseteq \{\xi \in \mathbb{R} : \xi > 0\} \subseteq t_2 \supseteq b_k\,. \tag{3.16}$$

In either case the definition (2.14) implies $0 \notin A - B$, and the assertion follows. □

THEOREM 3.13. *Let $A, B \in \mathbb{IB}$ for an admissible set of interval bounds $\mathbb{B}$. If $\{\xi \in \mathbb{R} : \xi \leq 0\} \notin \mathbb{B}$, then*

$$B \subseteq A/(A/B) \qquad \text{for all } A \neq \emptyset \text{ with } 0 \notin A \cup B \tag{3.17}$$

is true if and only if $\mathbb{B}$ is dense around 0.

Proof. If $\mathbb{B}$ is dense around 0, then by $0 \notin A \cup B$ and (3.11) we have $0 \notin A/B$, and the assertion follows by (3.2). If $\mathbb{B}$ is not dense around 0, then Theorem 3.9 implies that $0 \in A/B$ is possible although $0 \notin A \cup B$, so that $A/(A/B) = \texttt{NaI}$. □

4. Interval arithmetic based on floating-point endpoints. A generic choice of endpoints suitable for numerical computations is based on floating-point numbers. Let, for example, $\mathbb{F}$ denote the set of finite single (binary32) or double precision (binary64) floating-point numbers according to the IEEE 754 standard [1, 2]. Note that $\mathbb{F} = -\mathbb{F}$ and $0 \in \mathbb{F}$. Then $\{\{f\} : f \in \mathbb{F}\}$ is a weakly admissible set of interval bounds. As we have seen in the previous section, the mandatory inclusion monotonicity (3.2) is not satisfied because overflow is not taken care of.

We first give some examples that associativity of addition and multiplication as well as sub-distributivity is not satisfied. Denote the relative rounding error unit[2] by $\mathbf{u}$, so that $1-2\mathbf{u}, 1-\mathbf{u}, 1, 1+2\mathbf{u}, 1+4\mathbf{u}$ are consecutive floating-point numbers. Following we identify a floating-point number f with $\diamondsuit(f) = [\![\{f\}, \{f\}]\!]$ and use always interval operations. Then for $A = 1$, $B = 3\mathbf{u}$ and $C = -3\mathbf{u}$ we have

$$(A+B)+C = [1-\mathbf{u}, 1+2\mathbf{u}] \neq [1,1] = A+(B+C)\,, \tag{4.1}$$

[2] This is the maximal relative error of a floating-point operation; note that `eps` in Matlab gives $2\mathbf{u}$.

for $A = 1-2\mathbf{u}$ and $B = C = 1+2\mathbf{u}$ we have

$$A(BC) = [1, 1+4\mathbf{u}] \neq [1, 1+2\mathbf{u}] = (AB)C\ , \tag{4.2}$$

and for $A = 1-\mathbf{u}$ and $B = C = 1+2\mathbf{u}$,

$$(A+B)C = [2+4\mathbf{u}, 2+12\mathbf{u}] \not\subseteq [2+4\mathbf{u}, 2+8\mathbf{u}] = AC+BC\ . \tag{4.3}$$

The main reason that these fundamental arithmetic laws are not satisfied is that the distance of adjacent floating-point numbers decreases with their magnitude. As has been mentioned, for an equidistant number system such as fixed point numbers addition is exact as long as the result is in the representable range.

4.1. An admissible set of interval bounds not dense around 0 based on IEEE 754. Define

$$\texttt{realmin} := \min\{f : 0 < f \in \mathbb{F}\} \qquad \text{and} \qquad \texttt{realmax} := \max\{f : f \in \mathbb{F}\}\ . \tag{4.4}$$

A natural extension to an admissible set of interval bounds is to use

$$\mathbb{F}^* := \mathbb{F} \cup \{-\mathrm{H}, \mathrm{H}\} \qquad \text{with} \quad \mathrm{H} := \{\xi \in \mathbb{R} : \xi > \texttt{realmax}\}\ , \tag{4.5}$$

and to define $\mathbb{B} := \{\{f\} : f \in \mathbb{F}\} \cup \{-\mathrm{H}, \mathrm{H}\}$. Besides (3.1) and the inclusion monotonicity as in Theorem 3.3, not many mathematical properties are satisfied because $\mathbb{B}$ is not dense around 0: The inclusion of real numbers and operations as by (3.7) and (3.8) is not true, and of the remaining properties listed in Section 3 only the trivial "$\Leftarrow$" directions are satisfied. The equivalence in (3.15) is true if $\mathbb{F}$ includes gradual underflow because in this case $p-q=0$ is equivalent to $p=q$ for $p,q \in \mathbb{F}$. Practically speaking floating-point operations in the underflow range are rather slow; but banning un-normalized numbers spoils the equivalence in (3.15).

The interval operations can be realized using floating-point operations with directed rounding on the bounds of the input intervals. This corresponds to the usual approach, cf. [5, 7, 11]. Those operations $\circ_\nabla : \mathbb{F}^* \times \mathbb{F}^* \to \mathbb{F}^*$ and $\circ_\Delta : \mathbb{F}^* \times \mathbb{F}^* \to \mathbb{F}^*$ are defined except for division by zero by

$$p \circ_\nabla q := r \quad \text{and} \quad p \circ_\Delta q := s \qquad \text{where } [\![p,p]\!] \circ [\![q,q]\!] = [\![r,s]\!]\ , \tag{4.6}$$

identifying $f \in \mathbb{F}$ with $\{f\}$. If the additional quantities $\pm H$ are neither operands nor result and the operation is not division by zero, then the operations $\circ_\nabla$ and $\circ_\Delta$ are identical to the floating-point operations with directed rounding as defined in the IEEE 754 standard.

Provided $0 \notin [\![r,s]\!]$ in case of division and extending $\min_\preceq$ and $\max_\preceq$ in the obvious way for multiple arguments, one has

$$\begin{aligned} [\![p,q]\!] \circ [\![r,s]\!] \quad = \quad & \big[\!\big[\min\nolimits_\preceq \big(p \circ_\nabla r, p \circ_\nabla s, q \circ_\nabla r, q \circ_\nabla s\big), \\ & \quad \max\nolimits_\preceq \big(p \circ_\Delta r, p \circ_\Delta s, q \circ_\Delta r, q \circ_\Delta s\big) \big]\!\big]\ . \end{aligned} \tag{4.7}$$

In particular,

$$[\![p,q]\!] + [\![r,s]\!] = [\![p +_\nabla r, q +_\Delta s]\!] \quad \text{and} \quad [\![p,q]\!] - [\![r,s]\!] = [\![p -_\nabla s, q -_\Delta r]\!]\ . \tag{4.8}$$

For multiplication case distinctions can be used so that two multiplications are necessary unless both operands contain zero as an inner point, in which case four multiplications are necessary. For division always two divisions suffice, one in rounding downwards and one in rounding upwards.

TABLE 4.1
Addition in rounding upwards for $f \in \mathbb{F}$, $0 \geq \nu \in \mathbb{F}^*$ and $0 < \pi \in \mathbb{F}^*$

$p +_\Delta q$	$-\mathrm{H}$	ν	π	H
$-\mathrm{H}$	$-\mathrm{H}$	$-\mathrm{H}$	$\Delta(-\mathtt{realmax}+\pi)$	H
f	$\Delta(f-\mathtt{realmax})$	$\Delta(f+\nu)$	$\Delta(f+\pi)$	H
H	H	H	H	H

As usual, for $p,q \in \mathbb{F}$ the bounds can be computed directly using the directed rounding $\nabla : \mathbb{R} \to \mathbb{F}^*$ and $\Delta : \mathbb{R} \to \mathbb{F}^*$ defined by

$$\xi \in \mathbb{R} \quad \text{and} \quad \diamondsuit(\xi) = [\![p,q]\!] \qquad \Rightarrow \qquad \nabla(\xi) := p \quad \text{and} \quad \Delta(\xi) := q\,. \tag{4.9}$$

To simplify the exposition we identify $f \in \mathbb{F}$ with $\{f\} \in \mathbb{B}$ in the following. Then for the rounding upwards, for example, one has

$$\Delta(\xi) = \begin{cases} -\mathrm{H} & \text{if } \xi < -\mathtt{realmax} \\ \min\{f \in \mathbb{F} : \xi \leq f\} & \text{if } -\mathtt{realmax} \leq \xi \leq \mathtt{realmax} \\ \mathrm{H} & \text{if } \xi > \mathtt{realmax}\,. \end{cases} \tag{4.10}$$

Then, similar to IEEE 754, $p \circ_\nabla q := \nabla(p+q)$ and $p \circ_\Delta q := \Delta(p+q)$ for $p,q \in \mathbb{F}$, where $p+q \in \mathbb{R}$ denotes the real result of the sum. As an example, we show the results of $+_\Delta$ in Table 4.1.

Operations with directed rounding are defined in IEEE 754. By identifying H in $\mathbb{F}^*$ with ∞ in IEEE 754 those operations can be used except that $(-\infty) +_\Delta \infty = \mathtt{NaN}$ has to be taken care of.

4.2. An admissible set of interval bounds being dense around 0 based on IEEE 754. A natural extension to an admissible set of interval bounds being dense around 0 is

$$\mathbb{F}^* := \mathbb{F} \cup \{-\mathrm{H}, -\mathrm{T}, \mathrm{T}, \mathrm{H}\} \qquad \text{with } \mathrm{T} := \{\xi \in \mathbb{R} : 0 < \xi < \mathtt{realmin}\}\,, \tag{4.11}$$

and to define $\mathbb{B} := \{\{f\} : f \in \mathbb{F}\} \cup \{-\mathrm{H}, -\mathrm{T}, \mathrm{T}, \mathrm{H}\}$. Then, in contrast to the previous subsection, all mathematical properties listed in Section 3 are satisfied.

As before the interval operations can be realized using floating-point operations with directed rounding on the bounds of the input intervals as in (4.6) and (4.7).

For $p,q \in \mathbb{F}$ the bounds can again be computed directly using the directed rounding $\nabla : \mathbb{R} \to \mathbb{F}^*$ and $\Delta : \mathbb{R} \to \mathbb{F}^*$ defined in (4.9). For example, for $\mathbb{F}^*$ as in (4.11) and again identifying f with $\{f\}$, rounding upwards computes now as follows:

$$\Delta(\xi) = \begin{cases} -\mathrm{H} & \text{if } \xi < -\mathtt{realmax} \\ \min\{f \in \mathbb{F} : \xi \leq f\} & \text{if } -\mathtt{realmax} \leq \xi \leq -\mathtt{realmin} \\ -\mathrm{T} & \text{if } -\mathtt{realmin} < \xi < 0 \\ 0 & \text{if } \xi = 0 \\ \mathrm{T} & \text{if } 0 < \xi < \mathtt{realmin} \\ \min\{f \in \mathbb{F} : \xi \leq f\} & \text{if } \mathtt{realmin} \leq \xi \leq \mathtt{realmax} \\ \mathrm{H} & \text{if } \xi > \mathtt{realmax}\,. \end{cases} \tag{4.12}$$

Again we show as an example the results of $+_\Delta$ in Table 4.2. The remarks on multiplication and division in the previous subsection as well as (4.7) and (4.8) apply as before.

TABLE 4.2

Addition in rounding upwards for $v_1, v_2, \pi_1, \pi_2 \in \mathbb{F}$ *with* $-\mathtt{realmax} \le v_1, v_2 \le -\mathtt{realmin}$ *and* $\mathtt{realmin} \le \pi_1, \pi_2 \le \mathtt{realmax}$*; the lower half of the table is defined by symmetry*

$a +_\Delta b$	$-\mathrm{H}$	v_2	$-\mathrm{T}$	0	T	π_2	H
$-\mathrm{H}$	$-\mathrm{H}$	$-\mathrm{H}$	$-\mathrm{H}$	$-\mathrm{H}$	$\Delta(-\mathtt{realmax}+\mathtt{realmin})$	$\Delta(-\mathtt{realmax}+\pi_2)$	H
v_1		$\Delta(v_1+v_2)$	v_1	v_1	$\Delta(v_1+\mathtt{realmin})$	$\Delta(v_1+\pi_2)$	H
$-\mathrm{T}$			$-\mathrm{T}$	$-\mathrm{T}$	$\mathtt{realmin}$	π_2	H
0				0	T	π_2	H
T					$2\mathtt{realmin}$	$\Delta(\mathtt{realmin}+\pi_2)$	H
π_1						$\Delta(\pi_1+\pi_2)$	H
H							H

4.3. Examples and extensions. Let $\mathbb{B}$ be as in the previous subsection based on $\mathbb{F}^*$ in (4.11). We start with a few examples for the arithmetic described in the previous subsection. Let positive $p, q \in \mathbb{F}, p \le q$ be given such that $0 < p^2 < \mathtt{realmin}$ and $q^2 > \mathtt{realmax}$. For simplicity, we identify again $f \in \mathbb{F}$ with $\{f\} \in \mathbb{B}$. Then

$$C := [\![p,q]\!] \cdot [\![p,q]\!] = [\![\mathrm{T},\mathrm{H}]\!] \,. \tag{4.13}$$

Note that the range of C is $(0,\infty)$. Furthermore

$$D := [\![1,1]\!]/C = [\![\mathrm{T},\mathrm{H}]\!] = C \,. \tag{4.14}$$

Note that $0 \notin 1/[\![\mathrm{T},\mathrm{H}]\!]$. Moreover,

$$D := [\![1,1]\!]/[\![\mathrm{T},\mathrm{T}]\!] = [\![d,\mathrm{H}]\!] \,, \tag{4.15}$$

where $d := \diamondsuit(1/\mathtt{realmin})$. Furthermore,

$$\exp([\![-\mathrm{H},\mathrm{H}]\!]) = [\![\mathrm{T},\mathrm{H}]\!] \doteq (0,\infty) \quad \text{and} \quad \cosh([\![-\mathrm{H},\mathrm{H}]\!]) = [\![1,\mathrm{H}]\!] \doteq [1,\infty) \,, \tag{4.16}$$

or

$$\log([\![\mathrm{T},1]\!]) = [\![-\mathrm{H},0]\!] \doteq (-\infty,0] \qquad \text{and} \qquad \log([\![0,1]\!]) = \mathtt{NaI} \,. \tag{4.17}$$

In particular $\log(\exp([\![-\mathrm{H},\mathrm{H}]\!])) = [\![-\mathrm{H},\mathrm{H}]\!]$, $\exp(\log([\![\mathrm{T},1]\!])) = [\![\mathrm{T},1]\!]$ and

$$A \subseteq \log(\exp(A)) \qquad \text{for all } A \in \mathbb{IB} \,. \tag{4.18}$$

The "tiny" and "huge" quantities satisfy

$$\begin{aligned} &\diamondsuit(\xi) = [\![\mathrm{T},\mathrm{T}]\!] \;\Leftrightarrow\; 0 < \xi < \mathtt{realmin} \quad \text{and} \\ &\diamondsuit(\xi) = [\![\mathrm{H},\mathrm{H}]\!] \;\Leftrightarrow\; \xi > \mathtt{realmax} \,. \end{aligned} \tag{4.19}$$

Moreover, the interval arithmetic can be extended in several ways. For example, an additional quantity $\mathrm{E} := \{e\}$ may be added to $\mathbb{B}$, where $e \in \mathbb{R}$ denotes the base of the natural logarithm. The ordering (2.3) is clear from the definition. Then, for example,

$$\exp(\log([\![1,\mathrm{E}]\!])) = [\![1,\mathrm{E}]\!] \quad \text{and} \quad \log([\![\mathrm{E},\mathrm{E}]\!]) = [\![1,1]\!] \,. \tag{4.20}$$

Furthermore, the set $\mathbb{B}$ of bounds may be augmented to be dense around other real numbers. Denote by $\mathrm{pred}(p)$ and $\mathrm{succ}(p)$ the predecessor and successor of $p \in \mathbb{F}$, respectively. Define the new quantities

$$1^- := \{\xi \in \mathbb{R} : \mathrm{pred}(1) < \xi < 1\} \qquad \text{and} \qquad 1^+ := \{\xi \in \mathbb{R} : 1 < \xi < \mathrm{succ}(1)\}$$

and supplement $\mathbb{B}$ as after (4.11) by $\{\mathrm{pred}(1)\}, 1^-, \{1\}, 1^+, \{\mathrm{succ}(1)\}$. Then $\mathbb{B}$ is an admissible set of interval bounds being dense around 0 and 1, and, for example,

$$\tanh([\![0,30]\!]) = [\![0,1^-]\!], \quad 1 - [\![0,1^-]\!] = [\![T,1]\!], \quad \coth([\![\mathrm{T},50]\!]) = [\![1^+,\mathrm{H}]\!]\,. \tag{4.21}$$

Another example of an extension is to add quantities

$$\pi_2^- := \{\xi \in \mathbb{R} : p_1 < \xi < \pi/2\}, \quad \pi_2 := \{\pi/2\} \quad \text{and} \quad \pi_2^+ := \{\xi \in \mathbb{R} : \pi/2 < \xi < p_2\}$$

to $\mathbb{B}$, where $[p_1, p_2] \in \mathbb{IF}$ denotes the narrowest interval enclosing $\pi/2$. Then

$$\begin{aligned} &\tan([\![0,\pi_2^-]\!]) = [\![0,H]\!], \qquad &&\tan([\![0,\pi_2]\!]) = \mathtt{NaI}, \\ &\cos([\![\pi_2,\pi_2]\!]) = [\![0,0]\!], \qquad &&\mathrm{atan}([\![\mathrm{T},10^{20}]\!]) = [\![\mathrm{T},\pi_2^-]\!]\,. \end{aligned}$$

Arithmetic operations with such new quantities follow Definition 2.7. To save computing time, the narrowest enclosing interval may also be relaxed into a slightly wider one as a weak interval extension.

4.4. Realization based on IEEE 754. A floating-point format in IEEE 754 divides into the sign bit and some bits for the mantissa and for the exponent. The special quantities `Infinity` and `NaN` are represented by the maximal possible exponent with mantissa zero or nonzero, respectively. Thus a floating-point number is interpreted as `NaN` if its bit string contains the maximum possible exponent and at least one nonzero mantissa bit.

This gives a lot of freedom which is scarcely used, for example by discriminating between quiet and signalling `NaN`. In the shortest floating-point format binary32 there are 23 mantissa bits which leaves $2^{23} - 3$ additional possibilities besides the two mentioned ones. Therefore additional special quantities such as "tiny", Euler's constant E, quantities dense around 1 or $\pi/2$, or others as given in the previous subsection can be represented within the IEEE 754 floating-point formats, and also floating-point operations including directed rounding can be defined without jeopardizing the standard floating-point arithmetic.

5. An "ignoring input out of range" mode (iioor). In some applications, in particular in global optimization, it is advantageous to ignore inputs out of range. This is commonly known as "cset"-arithmetic and was proposed by Hansen and Walster, with theoretical foundation by Pryce, cf. [8]. It shifts a responsibility to the user to make sure that results are used in a proper way, see below. Our interval arithmetic defined in Section 2 can be altered in that sense, thus avoiding the result `NaI`.

More precisely, we use the set $\mathbb{IB}$ as intervals (rather than $\overline{\mathbb{IB}}$). Definition (2.14) is changed for $A, B \in \mathbb{IB}$ into

$$A \circ B := \bigcap \{C \in \mathbb{IB} : \alpha \circ \beta \in C \text{ for all } (\alpha,\beta) \in \mathrm{D}_\circ, \alpha \in A, \beta \in B\}\,, \tag{5.1}$$

where $\mathrm{D}_\circ$ is the range of definition of the operator $\circ$. The only difference to Definition (2.14) is that a zero in the denominator interval is ignored. Similarly, the Definition (2.19) of the natural interval extension $F : \mathbb{IB} \to \mathbb{IB}$ of a function $f : D_f \subseteq \mathbb{R} \to \mathbb{R}$ is replaced by

$$f(A) := \bigcap \{C \in \mathbb{IB} : f(\alpha) \in C \text{ for all } \alpha \in A \cap D_f\}\,. \tag{5.2}$$

If $A \cap D_f = \emptyset$, then $f(A) = \emptyset$. This implies, for example,

$$\log([\![0,1]\!]) = \log([\![-1,1]\!]) = [\![-\mathrm{H},0]\!] \doteq (-\infty,0] \,. \tag{5.3}$$

It is well-known that such an interval arithmetic is not suitable for the application of Brouwer's Fixed Point Theorem unless it is monitored that an input out of range occurred (for example by a flag).[3]

All properties in Section 3 remain valid *mutatis mutandis*, except Theorem 3.13: Here

$$1/(1/[\![0,1]\!]) = 1/[\![1,\mathrm{H}]\!] = [\![\mathrm{T},1]\!] \tag{5.4}$$

in the "iior"-mode of the arithmetic defined in Subsection 4.3, so that $A \subseteq 1/(1/A))$ is not necessarily true.

ACKNOWLEDGEMENTS. My thanks to Florian Bünger and Christian Jansson for comments on a preliminary version of the manuscript. In particular I am indebted to two anonymous referees for their most constructive and useful comments.

REFERENCES

[1] *ANSI/IEEE 754-1985: IEEE Standard for Binary Floating-Point Arithmetic*. New York, 1985.
[2] *ANSI/IEEE 754-2008: IEEE Standard for Floating-Point Arithmetic*. New York, 2008.
[3] J. Elstrodt. *Maß- und Integrationstheorie*. Springer- Verlag, Berlin Heidelberg, 1996.
[4] D.E. Knuth. *Surreal Numbers: How two ex-students turned on to pure mathematics and found total happiness*. Addison Wesley, Reading, Massachusetts, 1974.
[5] R.E. Moore. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, 1979.
[6] A. Neumaier. Vienna proposal for interval arithmetic.
[7] A. Neumaier. *Interval Methods for Systems of Equations*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1990.
[8] J. D. Pryce and G. F. Corliss. Interval Arithmetic with Containment Sets. *Computing*, 78:251–276, 2006.
[9] A. Robinson. *Non-standard analysis*. Princton University Press, 1996.
[10] S.M. Rump. INTLAB - INTerval LABoratory. In Tibor Csendes, editor, *Developments in Reliable Computing*, pages 77–104. Kluwer Academic Publishers, Dordrecht, 1999.
[11] S.M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010.

[3] For instance, $f(x) = \sqrt{x} - 1$ has no fixed point, but $F([\![-4,4]\!]) = [\![0,2]\!] \subseteq [\![-4,4]\!]$ in the "iioor"-mode.