Werk

Titel: An Axiomatic Approach to Rounded Computations.

Autor: KULISCH, U.

PURL: https://resolver.sub.uni-goettingen.de/purl?362160546_0018|log6

Kontakt/Contact

Digizeitschriften e.V.
SUB Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen

info@digizeitschriften.de

Numer. Math. 18, 1–17 (1971)
© by Springer-Verlag 1971

An Axiomatic Approach to Rounded Computations*

U. KULISCH

Herrn Lothar Collatz zum 60. Geburtstag gewidmet

Received January 9, 1970

Abstract. The present paper is intended to give an axiomatic approach to rounded computations. A rounding is defined as a monotone mapping of an ordered set into a subset, which in general is called a lower respectively an upper screen. The first chapter deals with roundings in ordered sets. In the second chapter further properties of roundings in linearly ordered sets are studied. The third chapter deals with the two most important applications, the approximation of the real arithmetic on a finite screen and the approximation of the real interval arithmetic on an upper screen. Beyond these examples various further applications are possible.

0. Introduction

It is a matter of fact that there hardly exist two computers of different companies which are using the same rounding procedure. The most frequently used programming languages like ALGOL-60, FORTRAN or PL/1 therefore do not make any assumption of special properties of the rounding procedure for floating point numbers. Numerical methods however, which are trying to get control of the rounding errors, have to make use of clearly defined rounding procedures. Therefore an axiomatic approach to rounded computations seems to be necessary.

There are already former attempts to define the approximation of the real arithmetic in a finite subset in an axiomatic way [6]. Further, more unconscious ideas in this direction can be found in [1–5]. We will try here to capture all the essential properties connected with rounded computations by setting up a system of axioms for more general problems, which permit a lot of applications.

There are many further problems in applied mathematics which nearly show the same behavior. Let for instance M be the set of all real $n \times n$ matrices and $\mathbb{P}(M)$ the power set of M. Then it is well known that M is an ordered set and we can define the set $I(M)$ of intervals over M. If we define a multiplication for elements $A, B \in I(M)$ by: $A \times B := \{a \times b | a \in A \wedge b \in B\}$, then it is easy to see, that $A \times B$ in general is an element of $\mathbb{P}(M)$ and not of $I(M)$. To get again an element of $I(M)$, we have to round the result $A \times B$ from $\mathbb{P}(M)$ into $I(M)$. This process is very similar to that of approximating the real arithmetic in a finite subsystem.

Therefore we shall define a rounding as an approximation of a set M with an order relation defined on M in a subset T of M or, more generally as an approximation of an algebraic structure $\{M, *\}$ $(*: M \times M \to M)$ in a subset T with an

* Sponsored by the Mathematics Research Center, Madison, Wisconsin, under Contract No.: DA-31-124-ARO-D-462.

approximating structure $\overline{*}$: $T \times T \to T$. It is quite clear, that this only can be done in a meaningful way if there is an order relation defined in M. To be clear, let us at first define the order relations, which we will use.

1. Roundings in Ordered Sets

Definition 1. Let M be a set with elements $a, b, c, \ldots$. The relation $\leqq$ in M is called an "order" or "order relation" and $\{M, \leqq\}$ is called an ordered set, if the axioms (O1), (O2) and (O3) are satisfied:

(O1) $\quad \bigwedge_a a \leqq a$ (reflexive)

(O2) $\quad \bigwedge_a \bigwedge_b \bigwedge_c \left((a \leqq b \wedge b \leqq c) \Rightarrow a \leqq c\right)$ (transitive)

(O3) $\quad \bigwedge_a \bigwedge_b \left((a \leqq b \wedge b \leqq a) \Rightarrow a = b\right)$ (antisymmetric).

An ordered set is called a "linearly ordered set", if in addition the property (O4) holds:

(O4) $\quad \bigwedge_a \bigwedge_b (a \leqq b \vee b \leqq a)$ (linearly ordered).

Definition 2. Let $\{M, \leqq\}$ be an ordered set and $T \subseteq M$. For all $a \in M$ we denote by $L_T(a) := \{b \mid b \in T \wedge b \leqq a\}$ respectively $U_T(a) := \{b \mid b \in T \wedge a \leqq b\}$ the set of all lower resp. upper bounds of a which are contained in T.

a) The subset $T \subseteq M$ is called a lower resp. an upper screen of M if

(S1) $\quad \bigwedge_{a \in M} L_T(a) \neq \emptyset \quad$ resp. $\quad \bigwedge_{a \in M} U_T(a) \neq \emptyset$

(S2) $\quad \bigwedge_{a \in M} \bigvee_{x \in L_T(a)} \bigwedge_{b \in L_T(a)} b \leqq x \quad$ resp. $\quad \bigwedge_{a \in M} \bigvee_{x \in U_T(a)} \bigwedge_{b \in U_T(a)} x \leqq b.$

We denote such an element x by $i\left(L_T(a)\right)$ resp. $o\left(U_T(a)\right)$; it is unique by (O3).

If $T \subseteq M$ is simultaneously a lower and an upper screen, then it is called a screen of M.

b) If T, S are lower resp. upper screens resp. screens of M, the property $T \subset S \subseteq M$ is expressed by saying T is coarser than S or S is finer than T.

Remark. If S, T are lower screens of M and $T \subset S \subseteq M$, then T is also a lower screen of S. The corresponding property holds for an upper screen and for a screen.

To illustrate Definition 2 let us consider some simple examples:

Example 1. Let $\mathbb{C}$ be the set of complex numbers, let

$$Z := \{\zeta \mid \zeta = \xi + i\eta \in \mathbb{C} \wedge |\xi| \leqq s \wedge |\eta| \leqq s\}, \quad \text{and let} \quad M := \mathbb{P}(Z)$$

be the powerset of Z. Then $\{\mathbb{P}(z), \subseteq\}$ is an ordered set. Let further $T \subseteq M$ be the set of all rectangles of Z with sides parallel to the real and imaginary axes. Then T is an upper screen of M.

Proof. (S1): $\bigwedge_{a \in M} \bigvee_{b \in T} a \subseteq b$, (Fig. 1 a).

(S2): We have to show that for all $a \in M$ the set $U_T(a) := \{b \mid b \in T \wedge a \subseteq b\}$ has a least element.

Each element $b \in T$ can be written in the form

$$b = \{\beta \mid \beta = \beta_1 + i\beta_2 \wedge \beta_1 \in [\beta_{11}; \beta_{12}] \wedge \beta_2 \in [\beta_{21}; \beta_{22}]\}.$$

Let $\alpha = \alpha_1 + i\alpha_2$. Since $a \in M$ is bounded, the limits $\inf\limits_{\alpha \in a} \alpha_1$, $\sup\limits_{\alpha \in a} \alpha_1$, $\inf\limits_{\alpha \in a} \alpha_2$, $\sup\limits_{\alpha \in a} \alpha_2$ exist. Let further

$$\bar{a} := \left\{\bar{\alpha} = \bar{\alpha}_1 + i\,\bar{\alpha}_2 \mid \bar{\alpha}_1 \in \left[\inf_{\alpha \in a} \alpha_1; \sup_{\alpha \in a} \alpha_1\right] \wedge \bar{\alpha}_2 \in \left[\inf_{\alpha \in a} \alpha_2; \sup_{\alpha \in a} \alpha_2\right]\right\}.$$

Then $a \subseteq \bar{a} \subseteq T$ and therefore $\bar{a} \in U_T(a)$. Further for all $b \in T$ with the property $a \subseteq b$ it follows $\bar{a} \subseteq b$, i.e. $\bar{a} = o\left(U_T(a)\right)$ (Fig. 1 b).

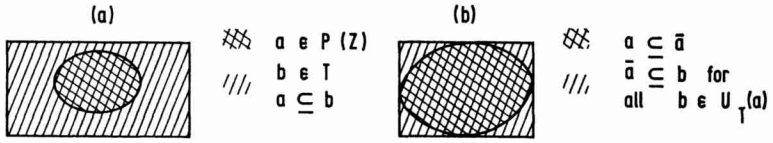


Fig. 1. Example for the concept of a screen

The set $L_T(a) := \{b \mid b \in T \wedge b \subseteq a\}$ obviously also satisfies the property (S1), (Fig. 2a), but not the property (S2) (Fig. 2b), i.e. T is not a lower screen of M and therefore no screen of M.

Fig. 2. Example for the concept of a screen

Example 2. Let $\hat{\mathbb{R}} := \{x \mid x \in \mathbb{R} \wedge |x| \leqq s\}$. Then $\{\hat{\mathbb{R}}, \leqq\}$ is a linearly ordered set. If $\hat{\mathbb{R}}_M$ is a finite subset of $\hat{\mathbb{R}}$ with the property $-s \in \hat{\mathbb{R}}_M$ and $s \in \hat{\mathbb{R}}_M$, then obviously $\hat{\mathbb{R}}_M$ is a screen of $\{\hat{\mathbb{R}}, \leqq\}$.

Example 3. Let now $[a_1; a_2]$ be a closed real interval, $I(\hat{\mathbb{R}}) = \{[a_1; a_2] \mid a_1, a_2 \in \hat{\mathbb{R}} \wedge a_1 \leqq a_2\}$, and $I(\hat{\mathbb{R}}_M) := \{[a_1, a_2] \mid a_1, a_2 \in \hat{\mathbb{R}}_M \wedge a_1 \leqq a_2\}$. Then $\{I(\hat{\mathbb{R}}), \subseteq\}$ is an ordered set and $I(\hat{\mathbb{R}}_M)$ is only an upper screen of $\{I(\hat{\mathbb{R}}), \subseteq\}$.

Definition 3. Let $\{M, \preceq\}$ be an ordered set and $T \subseteq M$ a lower resp. an upper screen of M resp. a screen of M. A mapping $\square : M \to T$ is called a rounding of M into T, if

(R1) $$\bigwedge_{a \in M} \bigwedge_{b \in M} (a \preceq b \Rightarrow \square a \preceq \square b) \quad \text{(monotone)}.$$

A rounding is called optimal, if

(R2) $$\bigwedge_{a \in T} \square a = a \quad \text{(optimal)}.$$

A rounding is called downward directed resp. upward directed if

(R3) $\quad \bigwedge_{a\in M} \square a \leqq a \quad$ resp. $\quad \bigwedge_{a\in M} a \leqq \square a$ (directed).

Lemma 1. Let $\{M, \leqq\}$ be an ordered set, $T \subseteq M$ a lower resp. an upper screen resp. a screen of M and $\square: M \to T$ a rounding of M into T. Then for all $a \in T$ the set $\square^{-1}a \subseteq M$ is convex.

Proof. Let $a_1, a_2 \in \square^{-1}a \subseteq M$ and $b \in M$ with the property $a_1 \leqq b \leqq a_2$. Then it follows from (R1) $\square a = \square a_1 \leqq \square b \leqq \square a_2 = \square a$ and hence from (O3) $\square b = \square a$, i.e. $b \in \square^{-1}a$.

Theorem 1. Let $\{M, \leqq\}$ be an ordered set, $T \subseteq M$ a lower resp. an upper screen resp. a screen of M and $\square: M \to T$ an optimal downward directed resp. upward directed rounding of M into T. Then

(R) $\quad \bigwedge_{a\in M} \square a = i\big(L_T(a)\big) \quad$ resp. $\quad \bigwedge_{a\in M} \square a = o\big(U_T(a)\big)$.

Further (R) $\Rightarrow$ [(R1), (R2), (R3)].

Proof. We show at first: (R) $\Rightarrow$ [(R1), (R2), (R3)]:

(R1): $\bigwedge_{a\in M} \bigwedge_{b\in M} \big(a \leqq b \Rightarrow L_T(a) \subseteq L_T(b) \Rightarrow \square a = i(L_T(a)) \leqq \square b = i(L_T(b))\big)$. It is easy to see that (R) $\Rightarrow$ (R2), (R3).

2. We show now: [(R1), (R2), (R3)] $\Rightarrow$ (R):

(R3) implies $\square a \in L_T(a)$ *. We show now that $\square a$ is a maximal element in $L_T(a)$. Let us assume that this assumption is not true. Then there exists an element $y \in L_T(a)$ with the property $\square a \prec y \leqq a$. If we apply the mapping $\square$ to this inequality, we get

$$\square(\square a) \underset{(R2)}{=} \square a \underset{(R1)}{\leqq} \square y \underset{(R1)}{\leqq} \square a \quad \text{and hence} \quad \square a \underset{(O3)}{=} \square y \underset{(R2)}{=} y.$$

This is a contradiction to the assumption $y \neq \square a$. Therefore $\square a$ is a maximal element of $L_T(a)$. Because of the fact that $L_T(a)$ has a greatest element (S2), the maximal element $\square a$ is equal to the unique greatest element $i\big(L_T(a)\big)$.

Corollary. There exists one and only one optimal downward directed resp. upward directed rounding of an ordered set $\{M, \leqq\}$ into a lower resp. an upper screen.

If $\{M, \leqq\}$ is an ordered set and T a lower resp. an upper screen of $\{M, \leqq\}$ then henceforth we will denote the uniquely determined optimal downward directed resp. optimal upward directed rounding of M into T by ∇ resp. $\triangle$. Then we obtain from Theorem 1:

(R) $\quad \bigwedge_{a\in M} \nabla a = i\big(L_T(a)\big) \quad$ resp. $\quad \bigwedge_{a\in M} \triangle a = o\big(U_T(a)\big)$.

Theorem 2. Let $\{M, \leqq\}$ be an ordered set and T and S both lower resp. both upper screens of M with the property $T \subseteq S \subseteq M$ **. Let further ∇, ∇_1, ∇_2 be

---

* We prove this and the following theorems only for the case of a lower screen. A corresponding proof holds for the upper screen.

** Let for instance T be the set of floating point numbers of a given compter and S the set of "long" floating point numbers.

the optimal downward directed roundings $\nabla\colon M\to T$, $\nabla_1\colon M\to S$, $\nabla_2\colon S\to T$ resp. Δ, Δ_1, Δ_2 the optimal upward directed roundings $\Delta\colon M\to T$, $\Delta_1\colon M\to S$, $\Delta_2\colon S\to T$, then

$$\bigwedge_{a\in M}\nabla a=\nabla_2(\nabla_1 a)\quad\text{resp.}\quad\bigwedge_{a\in M}\Delta a=\Delta_2(\Delta_1 a).$$

Proof. $T\subset S\Rightarrow L_T(a)\subset L_S(a)\Rightarrow\nabla a=i\big(L_T(a)\big)\leqq\nabla_1 a=i\big(L_S(a)\big)$. If we apply the mapping ∇_2 to this inequality, we get:

$$\nabla_2(\nabla a)\underset{(\mathrm{R}2)}{=}\nabla a\underset{(\mathrm{R}1)}{\leqq}\nabla_2(\nabla_1 a).\tag{1}$$

From (R3) we get $\nabla_2(\nabla_1 a)\leqq a$ and therefore

$$\nabla\big(\nabla_2(\nabla_1 a)\big)\underset{(\mathrm{R}2)}{=}\nabla_2(\nabla_1 a)\underset{(\mathrm{R}1)}{\leqq}\nabla a.\tag{2}$$

With (O3) we obtain from (1) and (2) $\nabla a=\nabla_2(\nabla_1 a)$.

We show by the simple

Example 4, that in general Theorem 2 is not valid for optimal, non-directed roundings of an ordered set into a screen. Let be $M:=[0;8]\subset\mathbb{R}$, $T:=\{0,4,8\}$ and $S:=\{0,1,2,3,4,5,6,7,8\}$ and let us define the roundings $\square\colon M\to T$, $\square_1\colon M\to S$ and $\square_2\colon S\to T$ as mappings to the nearest point of the screen with the further property, that the map of the middle of two neighborpoints of the screen is the lower neighbor. Then we get for $a=2.4\in M$: $\square_1 a=2$ and $\square_2(\square_1 a)=0\neq\square a=4$.

On some computers therefore the rounding is defined in the following way:

$$\square a:=\begin{cases}i\big(L_T(a)\big) & \text{for}\quad a-i\big(L_T(a)\big)<o\big(U_T(a)\big)-a\\ i\big(L_T(a)\big)\vee o\big(U_T(a)\big) & \text{for}\quad a-i\big(L_T(a)\big)=o\big(U_T(a)\big)-a\\ o\big(U_T(a)\big) & \text{for}\quad a-i\big(L_T(a)\big)>o\big(U_T(a)\big)-a,\end{cases}$$

where in the case $a-i\big(L_T(a)\big)=o\big(U_T(a)\big)-a$ the "$\vee$" is chosen by a random process. It is easy to see however, that Theorem 2 is not valid even in this case.

Definition 4. A non empty set M with an inner operation $*\colon M\times M\to M$ defined in M is called a "groupoid" and denoted by $\{M,*\}$. An element $e\in M$ with the property $\bigwedge_{a\in M}e*a=a$ respectively $\bigwedge_{a\in M}a*e=a$ is called a "left-" resp. a "right neutral element" of M. If e is simultaneously a left- and a right neutral element of M, then e is called a "neutral element" of M.

Example 5. $\{\mathbb{R},+\}$ and $\{\mathbb{P}(\mathbb{R}),+\}^\star$ are groupoids with the neutral elements 0 and $\{0\}$. $\{\mathbb{R},\times\}$ and $\{\mathbb{P}(\mathbb{R}),\times\}$ are groupoids with the neutral elements 1 and $\{1\}$. $\{\mathbb{R},-\}$ and $\{\mathbb{P}(\mathbb{R}),-\}$ are groupoids with the right neutral elements 0 and $\{0\}$. $\{\mathbb{R},/\}$ and $\{\mathbb{P}(\mathbb{R}),/\}$ are groupoids with the right neutral elements 1 and $\{1\}$.

Definition 5. Let $\{M,*\}$ be a groupoid and $\{M,\leqq\}$ an ordered set. Let further T be a lower resp. an upper screen of $\{M,\leqq\}$. If in T an inner operation

---

* With $\mathbb{P}(M)$ we denote the power set of a set M.

$\overline{*}: T \times T \to T$ is defined, the groupoid $\{T, \overline{*}\}$ is called a "screen groupoid" of $\{M, *\}$ with respect to the order relation $\leqq$, if

$$\text{(RG1)} \qquad \bigwedge_{a,b,c,d \in T} (a*b \leqq c*d \Rightarrow a \overline{*} b \leqq c \overline{*} d).$$

A screen groupoid is called optimal if

$$\text{(RG2)} \qquad \bigwedge_{a \in T} \bigwedge_{b \in T} \big((a*b \in T) \Rightarrow a \overline{*} b = a*b\big).$$

A screen groupoid is called a lower resp. an upper screen groupoid, if

$$\text{(RG3)} \qquad \bigwedge_{a \in T} \bigwedge_{b \in T} a \overline{*} b \leqq a*b \quad \text{resp.} \quad \bigwedge_{a \in T} \bigwedge_{b \in T} a*b \leqq a \overline{*} b.$$

Theorem 3. Let $\{M, *\}$ be a groupoid, $\{M, \leqq\}$ an ordered set, T a lower resp. an upper screen resp. a screen of $\{M, \leqq\}$ and $\square: M \to T$ a mapping of M into T. If we define in T an inner operation $\overline{*}: T \times T \to T$ by $\bigwedge_{a \in T} \bigwedge_{b \in T} a \overline{*} b := \square (a*b)$, the following properties hold:

1. (R1) $\Rightarrow$ (RG1).
2. (R2) $\Rightarrow$ (RG2).
3. (R3) $\Rightarrow$ (RG3).
4. If $\{M, *\}$ is commutative, $\{T, \overline{*}\}$ is also commutative.

Proof. 1. For all $a, b, c, d \in T$ with the property $a*b \leqq c*d$ we get from (R1): $a \overline{*} b = \square (a*b) \leqq \square (c*d) = c \overline{*} d$.

2. For all $a, b \in T$ with the property $a*b \in T$ we get from (R2): $a \overline{*} b = \square (a*b) = a*b$.

3. For all $a, b \in T$ we get from (R3): $a \overline{*} b = \square (a*b) \leqq a*b$.

4. For all $a, b \in T$ we have $a \overline{*} b = \square (a*b) = \square (b*a) = b \overline{*} a$.

Lemma 2. Let $\{M, *\}$ be a groupoid with a right neutral element e, $\{M, \leqq\}$ an ordered set, T a lower resp. an upper screen resp. a screen of $\{M, \leqq\}$ with the property $e \in T$ and $\{T, \overline{*}\}$ an optimal screen groupoid. Then e is also a right neutral element of $\{T, \overline{*}\}$.

Proof. $\bigwedge_{a \in T} (a*e \in T) \underset{\text{(RG 2)}}{\Rightarrow} \bigwedge_{a \in T} (a \overline{*} e = a*e = a)$.

Theorem 4. Let $\{M, *\}$ be a groupoid with a right neutral element e, $\{M, \leqq\}$ an ordered set, T a lower resp. an upper screen resp. a screen of $\{M, \leqq\}$ with the property $e \in T$ and $\nabla: M \to T$ resp. $\Delta: M \to T$ the optimal directed roundings of M into T. Let further $\{T, \overline{*}\}$ be an optimal lower resp. upper screen groupoid of $\{M, *\}$ with respect to the order relation $\leqq$. Then

$$\text{(RG)} \qquad \bigwedge_{a \in T} \bigwedge_{b \in T} a \overline{*} b = \nabla (a*b) \quad \text{resp.} \quad \bigwedge_{a \in T} \bigwedge_{b \in T} a \overline{*} b = \Delta (a*b).$$

Proof. Let $a, b \in T$ and $x := \nabla (a*b)$. Then we get from (R3) $x = x*e = \nabla (a*b) \leqq a*b$ and hence with (RG1) and the above lemma

$$x \overline{*} e = x = \nabla (a*b) \leqq a \overline{*} b. \tag{3}$$

From (RG3) $a \overline{*} b \leqq a * b$ we get with (R2) and (R1)

$$\nabla(a \overline{*} b) = a \overline{*} b \leqq \nabla(a * b). \tag{4}$$

Because of (O3) we get from (3) and (4) $a \overline{*} b = \nabla(a * b)$.

From the Theorems 1, 3 and 4 we obtain immediately the

Corollary. Let $\{M, *\}$ be a groupoid with a right neutral element e, $\{M, \leqq\}$ an ordered set, T a lower resp. an upper screen resp. a screen of $\{M, \leqq\}$ with the property $e \in T$. Then there exists one and only one optimal lower resp. upper screen groupoid of $\{M, *\}$ with respect to the order relation $\leqq$.

If the assumptions of this corollary are fulfilled we shall denote henceforth the uniquely determined optimal lower resp. upper screen-groupoid of $\{M, *\}$ by $\{T, \underset{\vee}{*}\}$ resp. $\{T, \hat{*}\}$. With the Theorems 3 and 4 we have then

$$\text{(RG)} \qquad \bigwedge_{a \in T} \bigwedge_{b \in T} a \underset{\vee}{*} b = \nabla(a * b) \quad \text{resp.} \quad \bigwedge_{a \in T} \bigwedge_{b \in T} a \hat{*} b = \Delta(a * b).$$

Example 6. Let again $Z := \{\zeta \mid \zeta = \xi + i\eta \in \mathbb{C} \wedge |\xi| \leqq s \wedge |\eta| \leqq s\}$, $M := \mathbb{P}(Z)$ and $T \subset M$ the set of all rectangles of Z with sides parallel to the real and imaginary axis. Then, as we know from Example 1, T is an upper screen of $\{M, \subseteq\}$. Let us now define a multiplication $\times : M \times M \to M$ by

$$\bigwedge_{a \in M} \bigwedge_{b \in M} a \times b := \{\alpha \times \beta \mid \alpha, \beta \in Z \wedge \alpha \in a \wedge \beta \in b\}.$$

Then obviously the element $\{1\}$ is a neutral element of this multiplication, i.e. $\{M, \times\}$ is a groupoid with neutral element $\{1\}$. Let us consider now the special elements $a, b \in M$ (Fig. 3 a):

$$a := \{\zeta \mid \zeta = \xi + i\eta \wedge \xi \in [1; 2] \wedge \eta \in [0; 1]\}, \quad b = \{e^{\frac{i\pi}{4}}\} = \{\tfrac{1}{2}\sqrt{2}\,(1+i)\}.$$

a and b are also elements of $T \subset M$. Fig. 3 b shows the product $a \times b$ and the set $a \hat{\times} b = \Delta(a \times b)$.

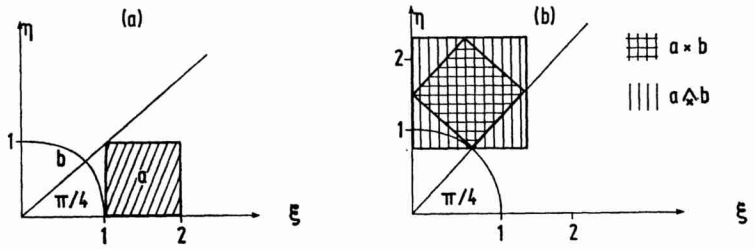


Fig. 3. Multiplication in a screengroupoid

Definition 5. A non-empty set M is called an "ordered groupoid" and denoted by $\{M, *, \leqq\}$, if the following axioms are satisfied

(G1) M is a groupoid.

(G2) M is an ordered set.

(G3) The algebraic structure of M is compatible (isotone) with the order structure:

$$a \leqq b \Rightarrow \left(\bigwedge_c a*c \leqq b*c \wedge c*a \leqq c*b\right).$$

Remark. The compatibility condition (G3) is equivalent to the condition:

(G3') for all $a, b, c, d \in M$ with the property $a \leqq c \wedge b \leqq d \Rightarrow a*b \leqq c*d$, since (G3) $\Rightarrow a*b \leqq a*d \leqq c*d$ and (G3') $\Rightarrow$ (G3). We shall apply the compatibility condition in general in the form (G3').

Example 7. $\{\mathbb{R}, +, \leqq\}$ is an ordered groupoid, $\{\mathbb{R}, \times\}$ is a groupoid, $\{\mathbb{R}, \leqq\}$ is an ordered set, but $\mathbb{R}$ is no ordered groupoid with respect to the multiplication and the order relation $\leqq$. With $\mathbb{R}_+ := \{x \mid x \in \mathbb{R} \wedge x \geqq 0\}$ however $\{\mathbb{R}_+, \times, \leqq\}$ is an ordered groupoid. Further it is well known from interval analysis, that for all $* \in \{+, -, \times, /\}$ $\{\mathbb{IP}(\mathbb{R}), *, \subseteq\}$ is an ordered groupoid.

Lemma 3. Let $\{M, *, \leqq\}$ be an ordered groupoid and $\{T, \overline{*}\}$ a screen groupoid of $\{M, *, \leqq\}$. Then $\{T, \overline{*}, \leqq\}$ is also an ordered groupoid.

Proof. From (G3') in $\{M, *, \leqq\}$ and (RG1) follows (G3') in $\{T, \overline{*}, \leqq\}$.

Theorem. 5. Let $\{M, *, \leqq\}$ be an ordered groupoid, T a lower resp. an upper screen resp. a screen of $\{M, *, \leqq\}$ and $\square: M \to T$ a downward directed resp. upward directed rounding of M into T. If in T an inner operation $\overline{*}: T \times T \to T$ is defined by

$$a \overline{*} b := \square(a*b), \tag{5}$$

then for all $a, b \in M$ the following inequalities are satisfied:

$$(\square a) \overline{*} (\square b) \leqq \square(a*b) \leqq a*b \quad \text{resp.} \quad a*b \leqq \square(a*b) \leqq (\square a) \overline{*} (\square b). \tag{6}$$

Proof. The inequality $\square(a*b) \leqq a*b$ resp. $a*b \leqq \square(a*b)$ is an immediate consequence of (R3). In the case of a downward directed rounding we get further from (R3) $\square a \leqq a$ and $\square b \leqq b$ and therefore from (G3') $\square a * \square b \leqq a*b$. If we apply the mapping $\square$ to this inequality, (6) follows from (RG1) and (5).

A particular case of Theorem 5 is the

Corollary. Let $\{M, *, \leqq\}$ be an ordered groupoid, $\{T, \underset{\vee}{*}, \leqq\}$ an optimal lower screen groupoid and $\{T, \hat{*}, \leqq\}$ an optimal upper screen groupoid of $\{M, *, \leqq\}$ then for all $a, b \in M$ the following inequalities are satisfied:

$$(\nabla a) \underset{\vee}{*} (\nabla b) \leqq \nabla(a*b) \leqq a*b \quad \text{resp.} \quad a*b \leqq \Delta(a*b) \leqq (\Delta a) \hat{*} (\Delta b).$$

Remark 1. The inequality $(\square a) \overline{*} (\square b) \leqq \square(a*b)$ resp. $\square(a*b) \leqq (\square a) \overline{*} (\square b)$ in (6) tells us, that the lower resp. upper screen groupoid (5) in general is no homomorphic image of the ordered groupoid $\{M, *, \leqq\}$. In concrete cases we shall show by simple examples, that in general the inequality sign holds. The equality sign does not even hold in the case of an optimal lower or upper screen groupoid.

Remark 2. The outer inequality (6):

$$(\square a) \overline{*} (\square b) \leqq a*b \quad \text{resp.} \quad a*b \leqq (\square a) \overline{*} (\square b) \tag{6'}$$

immediately carries over to expressions, which are composed by more than one operation. That means, that the result of any computation in a lower resp. an upper screen groupoid (5) will be less resp. greater than the result of the corresponding calculation in M. This is also true in an optimal lower resp. upper screen groupoid.

If in M more than one operation is defined and all these operations get approximated in T by lower resp. upper screen groupoids (5), then the above statement is valid also in expressions which are composed by different operations.

Because of Theorem 5 it is possible for instance to approximate the interval arithmetic on a computer in a non optimal, but meaningful way if the special implementation of the rounding procedure for floating point numbers on the computer makes it impossible to carry out interval calculations with optimal roundings.

Theorem 6. Let $\{M, *, \leqq\}$ be an ordered groupoid with a right neutral element e, T and S both lower resp. both upper screens of $\{M, *, \leqq\}$ with the property $e \in T$ and S finer than T, i.e. $T \subset S \subseteq M$. Let further $\nabla: M \to T$, $\nabla_1: M \to S$, $\nabla_2: S \to T$ be the optimal downward directed roundings resp. $\Delta: M \to T$, $\Delta_1: M \to S$, $\Delta_2: S \to T$ the optimal upward directed roundings and $\{T, \underset{\vee}{*}, \leqq\}$, $\{S, \underset{\vee}{*}_1, \leqq\}$, $\{T, \underset{\vee}{*}_2, \leqq\}$ resp. $\{T, \hat{*}, \leqq\}$, $\{S, \hat{*}_1, \leqq\}$, $\{T, \hat{*}_2, \leqq\}$ the corresponding optimal lower resp. upper screen groupoids. Then

a) $\{T, \underset{\vee}{*}, \leqq\} = \{T, \underset{\vee}{*}_2, \leqq\}$ resp. $\{T, \hat{*}, \leqq\} = \{T, \hat{*}_2, \leqq\}$.

b) $\bigwedge_{a \in M} \bigwedge_{b \in M} (\nabla a) \underset{\vee}{*} (\nabla b) \leqq (\nabla_1 a) \underset{\vee}{*}_1 (\nabla_1 b) \leqq a * b$

resp. $\bigwedge_{a \in M} \bigwedge_{b \in M} a * b \leqq (\Delta_1 a) \hat{*}_1 (\Delta_1 b) \leqq (\Delta a) \hat{*} (\Delta b)$,

i.e. a finer optimal lower resp. upper screen groupoid always leads to no worse results.

Proof. a) For all $a, b \in T$ we get from Theorem 4: $a \underset{\vee}{*}_2 b = \nabla_2 (a \underset{\vee}{*}_1 b)$ and since from Theorem 2 $a \underset{\vee}{*} b = \nabla (a * b) = \nabla_2 \big(\nabla_1 (a * b)\big) = \nabla_2 (a \underset{\vee}{*}_1 b) = a \underset{\vee}{*}_2 b$.

b) $(\nabla a) \underset{\vee}{*} (\nabla b) = \nabla \big((\nabla a) * (\nabla b)\big) = \nabla_2 \big(\nabla_1 ((\nabla a) * (\nabla b))\big) = \nabla_2 \big((\nabla a) \underset{\vee}{*}_1 (\nabla b)\big)$

$\underset{(R3)}{\leqq} (\nabla a) \underset{\vee}{*}_1 (\nabla b) = \nabla_2 (\nabla_1 a) \underset{\vee}{*}_1 \nabla_2 (\nabla_1 b) \underset{(G3')}{\leqq} (\nabla_1 a) \underset{\vee}{*}_1 (\nabla_1 b)$.

Theorem 6 involves some possibilities to check the influence of the rounding errors in numerical computations.

2. Roundings in Linearly Ordered Sets

In the previous chapter we have seen that in ordered sets the optimal directed roundings are distinguished by some important properties (Theorems 1, 2, 3, 4, 6). Furthermore in linearly ordered sets we shall get some further important properties for optimal non-directed roundings.

Lemma 4. Let $\{M, \leqq\}$ be a linearly ordered set, T a screen of M and $\nabla: M \to T$ resp. $\Delta: M \to T$ the optimal downward directed resp. upward directed rounding of M into T. Then for all $a \in M$ there does not exist an element $b \in T$ with the property $\nabla a \prec b \prec \Delta a$.

Proof. We know from Theorem 1, that $\nabla a = i(L_T(a))$ and $\Delta a = o(U_T(a))$. Let us assume now, that the lemma would not be true, then from (O4) we get $[i(L_T(a)) \prec b \leqq a$ or $a \leqq b \prec o(U_T(a))]$, which is in contradiction to the definition of $i(L_T(a))$ or $o(U_T(a))$.

Theorem 7. Let $\{M, \leqq\}$ be a linearly ordered set, T a screen of M, $\nabla: M \to T$ resp. $\Delta: M \to T$ the optimal downward directed resp. upward directed rounding of M into T and $\square: M \to T$ any further optimal rounding of M into T. Then, because of (R2), for all $a \in T \subseteq M$ we have $\square a = \nabla a = \Delta a = a$. For all $a \in M$ with the property $a \notin T$ we can split up the interval $I := [\nabla a;\ \Delta a] \subseteq M$ into subsets $I_1, I_2 \subseteq I$ such, that

$$\square a = \begin{cases} \nabla a & \text{for all} \quad a \in I_1 \\ \Delta a & \text{for all} \quad a \in I_2 \end{cases} \quad \text{with} \quad I_1 \prec I_2 \text{*} \quad \text{and} \quad I_1 \cup I_2 = I.$$

Proof. We know from Theorem 1, that $\nabla a = i(L_T(a)) \leqq a \leqq \Delta a = o(U_T(a))$. If we apply the mapping $\square$ to this inequality we get

$$\square(\nabla a) \underset{(R2)}{=} \nabla a \underset{(R1)}{\leqq} \square a \underset{(R1)}{\leqq} \square(\Delta a) \underset{(R2)}{=} \Delta a.$$

Therefore with the above Lemma 4 we have $\square a = (\nabla a$ or $\Delta a)$. Let now $I_1 := \{x \mid x \in I \wedge \square x = \nabla a\}$ and $I_2 := \{x \mid x \in I \wedge \square x = \Delta a\}$. For all $x \in I$ we have $\nabla a \leqq x \leqq \Delta a$. Hence (R1) implies $\nabla a \leqq \square x \leqq \Delta a$ and we get with Lemma 4 $I_1 \cup I_2 = I$. Further $I_1 \prec I_2$; let us assume otherwise, it would be $i_2 \leqq i_1$ ((O4)) for some pair $i_1 \in I_1$ and $i_2 \in I_2$, then (R1) would imply $\square i_2 \leqq \square i_1$, which is in contradiction to the assumption $\square i_1 = \nabla a \prec \square i_2 = \Delta a$ $(a \notin T!)$.

Remark. It follows from Theorem 7, that every optimal rounding of a linearly ordered set into a screen can be expressed by the optimal directed roundings ∇a and Δa. Different optimal roundings distingush each other in an interval $I := [a_1; a_2]$ with neighboring screenpoints a_1, a_2 only by the way of splitting I into subsets $I_1 \prec I_2$. The limits are the optimal directed roundings

$$\nabla a\ (I_1 = I - \{a_2\},\ I_2 = \{a_2\}) \quad \text{and} \quad \Delta a\ (I_1 = \{a_1\},\ I_2 = I - \{a_1\}).$$

Theorem 8. Let $\{M, *\}$ be a groupoid with a right neutral element e, $\{M, \leqq\}$ a linearly ordered set, T a lower resp. an upper screen resp. a screen of $\{M, \leqq\}$ with the property $e \in T$ and $\nabla: M \to T$ resp. $\Delta: M \to T$ the optimal downward resp. upward directed rounding of M into T. Let further be $\{T, \underset{\vee}{*}\}$ resp. $\{T, \hat{*}\}$ the optimal lower resp. upper screen groupoid and $\{T, \bar{*}\}$ any further optimal screen groupoid of $\{M, *\}$ with respect to the linear order relation $\leqq$. Then because of (RG2), for all $a, b \in T$ with the property $a * b \in T$ we have $a \bar{*} b = a \underset{\vee}{*} b = a \hat{*} b = a * b$. For all $a, b \in T$ with the property $a * b \notin T$ we get

$$a \bar{*} b = \begin{cases} a \underset{\vee}{*} b & \text{for} \quad a * b \in I_1 \\ a \hat{*} b & \text{for} \quad a * b \in I_2 \end{cases} \quad \text{with} \quad I_1 \prec I_2 \quad \text{and} \quad I_1 \cup I_2 = I$$

and $I^* := [\nabla(a * b);\ \Delta(a * b)] \subseteq M$, $N := \{x * y \mid x, y \in T\} \subseteq M$ and $I := I^* \cap N$, $I_1, I_2 \subseteq I$.

* For $U, V \subseteq M$ we write $U \leqq V$, if $u \leqq v$ for all $u \in U$ and all $v \in V$.

Further there exists at least one optimal rounding $\square: M \to T$ with the property $a \overline{*} b = \square(a * b)$ for all $a, b \in T$.

Proof. We known from Theorem 4 that $a \underset{\vee}{*} b = \nabla(a * b)$ and $a \hat{*} b = \Delta(a * b)$. Let now be $I_1 := \{x * y \mid x, y \in T,\ x * y \in I \wedge x \overline{*} y = a \underset{\vee}{*} b\}$ and $I_2 := \{x * y \mid x, y \in T,\ x * y \in I \wedge x \overline{*} y = a \hat{*} b\}$. For all $x, y \in T$ with the property $x * y \in I$ we have $a \underset{\vee}{*} b = \nabla(a * b) \leqq x * y \leqq \Delta(a * b) = a \hat{*} b$, respectively $(a \underset{\vee}{*} b) * e \leqq x * y \leqq (a \hat{*} b) * e$. Because of Lemma 2 and Theorems 3 and 4 we get

$$(a \underset{\vee}{*} b) \overline{*} e = a \underset{\vee}{*} b = \nabla(a * b) \underset{(\mathrm{RG}\,1)}{\leqq} x \overline{*} y \underset{(\mathrm{RG}\,1)}{\leqq} (a \hat{*} b) \overline{*} e = a \hat{*} b = \Delta(a * b). \tag{7}$$

Substituting $x = a$ and $y = b$ in (7) we get with Lemma 4 $a \overline{*} b = (a \underset{\vee}{*} b$ or $a \hat{*} b)$. From (7) we get also for all $x, y \in T$ with the property $x * y \in I$, $x \overline{*} y = (a \underset{\vee}{*} b$ or $a \hat{*} b)$, i.e. $I_1 \cup I_2 = I$. Further the inequality $I_1 \prec I_2$ holds: let us assume otherwise, that we would have $a_2 * b_2 \leqq a_1 * b_1$ ((O4)) with $a_1, b_1, a_2, b_2 \in T$ and $a_1 * b_1 \in I_1$ and $a_2 * b_2 \in I_2$; then (RG1) would imply $a_2 \overline{*} b_2 \leqq a_1 \overline{*} b_1$, which is in contradiction to the assumption $a_1 \overline{*} b_1 = \nabla(a * b) \leqq a_2 \overline{*} b_2 = \Delta(a * b)$, $(a * b \notin T)$.

To prove the last property of Theorem 8, let us call a pair of sets (I_1^*, I_2^*) an extension of the pair (I_1, I_2), if the following properties hold: $I_1 \subseteq I_1^*$, $I_2 \subseteq I_2^*$, $I_1^* \prec I_2^*$ and $I_1^* \cup I_2^* = I^*$. Clearly in general there exists more than one extension (I_1^*, I_2^*) of (I_1, I_2). Let now (I_1^*, I_2^*) be an extension of (I_1, I_2) and let us define a mapping $\square: M \to T$ with the property: for all $a \in T \subseteq M$ let $\square a = a$ and for all $a \in M$ with the property $a \notin T$ let be

$$\square a = \begin{cases} \nabla a, & \text{for all } a \in I_1^* \\ \Delta a, & \text{for all } a \in I_2^*. \end{cases}$$

Then obviously $\square$ is an optimal rounding of M into T and for all $a, b \in T$ with the property $a * b \in I$ we have $a \overline{*} b = \square(a * b)$. This completes the proof of Theorem 8.

Remarks. Let $\{M, *\}$ be a groupoid with a right neutral element e and $\leqq$ a linear order relation in M. Then it follows from Theorem 8 that the result of any operation in an optimal screen groupoid can be expressed by the appropriate results in the optimal lower or the optimal upper screen groupoid. In an interval $I^* := [a_1; a_2]$ with neighboring screenpoints a_1, a_2 different optimal screen groupoids distinguish each other only by the way of splitting $I := I^* \cap N$ into subsets $I_1 \prec I_2$. The limits are the optimal lower screen groupoid respectively the optimal upper screen groupoid.

Further it is worth remarking, that from Theorems 3, 4, 7 and 8 follows: the result of an operation $a \overline{*} b$ of any optimal screen groupoid $\{T, \overline{*}\}$ can be expressed by the operation $a * b$ in $\{M, *\}$ and the optimal downward directed and the optimal upward directed rounding of M into T:

$$a \overline{*} b = \begin{cases} \nabla(a * b) & \text{for all } a * b \in I_1 \\ \Delta(a * b) & \text{for all } a * b \in I_2 \end{cases} \quad \text{with } I_1 \prec I_2 \text{ and } I_1 \cup I_2 = I.$$

It is easy to show, by simple examples, that the Theorems 7 and 8 in general are not valid in ordered sets.

3. Applications

Let us now consider two applications, the approximation of the real arithmetic and the approximation of the real interval-arithmetic on a given computer.

It is a characteristic feature of all computing machines that the range of all variables is bounded. For instance with a planimeter it is only possible to measure surfaces which lie within the reach of its arms. The real numbers representable on a given computer in general cover a very large region, but in principle we have just the same situation. If the range of numbers is not sufficient to solve a given problem, the problem must be scaled in a suitable way. In practice however it is convenient and usual at first to start the computation. Most of the computers interrupt the computation with an error announcement if the available range of numbers is exceeded. Then the problem has to be scaled. Nevertheless the scaling belongs to the preparation of the proper computation.

Without loss of generality we may assume therefore, that all considerations of this chapter deal with a bounded set of real numbers. We will denote this set by

$$\hat{\mathbb{R}} := \{x \mid x \in \mathbb{R} \wedge |x| \leqq s\}.$$

3.1 Approximation of the Real Arithmetic in a Finite Subsystem

The set $\{\hat{\mathbb{R}}, \leqq\}$ is a linearly ordered set. Let $\hat{\mathbb{R}}_M$ be a finite subset of $\hat{\mathbb{R}}$ with the property $-s \in \hat{\mathbb{R}}_M$ and $s \in \hat{\mathbb{R}}_M$. Then obviously $\hat{\mathbb{R}}_M$ is a screen of $\{\hat{\mathbb{R}}, \leqq\}$. For instance $\hat{\mathbb{R}}_M$ may be the set of real numbers representable on a given computer. Because of this application let us assume a further property of $\hat{\mathbb{R}}_M$:

$$\text{(S3)} \qquad \bigwedge_{a \in \hat{\mathbb{R}}_M} -a \in \hat{\mathbb{R}}_M.$$

We know from Theorem 1 that we can define the optimal downward directed resp. the optimal upward directed roundings $\nabla: \hat{\mathbb{R}} \to \hat{\mathbb{R}}_M$ resp. $\Delta: \hat{\mathbb{R}} \to \hat{\mathbb{R}}_M$ by

$$\bigwedge_{a \in \hat{\mathbb{R}}} \nabla a := i\left(L_{\hat{\mathbb{R}}_M}(a)\right) \quad \text{resp.} \quad \bigwedge_{a \in \hat{\mathbb{R}}} \Delta a := o\left(U_{\hat{\mathbb{R}}_M}(a)\right). \tag{8}$$

Because of Theorem 7 every further optimal rounding $\square: \hat{\mathbb{R}} \to \hat{\mathbb{R}}_M$ can be expressed by the optimal downward directed and the optimal upward directed roundings (8).

In practice the following optimal roundings are often used:

$$\square_1 a := \begin{cases} \nabla a, & \text{if} \quad a \leqq (\nabla a + \Delta a)/2 \\ \Delta a, & \text{if} \quad a > (\nabla a + \Delta a)/2, \end{cases} \tag{9}$$

$$\square_2 a := \begin{cases} \nabla a, & \text{if} \quad a > 0 \\ \Delta a, & \text{if} \quad a < 0, \end{cases} \tag{10}$$

$$\square_3 a := \begin{cases} \nabla a, & \text{if} \quad a < 0 \\ \Delta a, & \text{if} \quad a > 0. \end{cases} \tag{11}$$

The roundings ∇, Δ and every further optimal rounding $\square$ satisfy the following properties, which in the cases of ∇ and Δ even define them:

(R1) $\bigwedge_{a\in\hat{\mathbb{R}}}\bigwedge_{b\in\hat{\mathbb{R}}}(a\leqq b\Rightarrow\nabla a\leqq\nabla b \quad \text{resp.} \quad \Delta a\leqq\Delta b \quad \text{resp.} \quad \square a\leqq\square b)$.

(R2) $$\bigwedge_{a\in\hat{\mathbb{R}}_M}\nabla a=\Delta a=\square a=a.$$

(R3) $$\bigwedge_{a\in\hat{\mathbb{R}}}\bigwedge_{b\in\hat{\mathbb{R}}}(\nabla a\leqq a \quad \text{resp.} \quad a\leqq\Delta a).$$

Because of (S3) the roundings Δ and ∇ are not independent of each other. For instance we can express the rounding Δ by the rounding ∇:

$$\Delta a=-\nabla(-a). \tag{12}$$

From (12) and Theorem 7 we get as an important result that the realization of one of the optimal, directed roundings ∇ or Δ on a computer is already sufficient to carry out every other optimal rounding on this computer.

The question, which of the roundings ∇ or Δ should be realized on a computer, can be answered by ∇. It is easy to see that we get ∇a for $a>0$ if we truncate the last digits. If we describe this process by $t(a)$ (truncation), we have $\nabla a=t(a)$ for $a>0$. For $a<0$ we can get ∇a just as easily, if we express negative numbers on the computer by their twos—or tens complement. We have then $\nabla a=t(x-a)-x$ with a suitable chosen x such, that $x-a$ is the twos or tens complement of a. This rounding procedure can be realized as easily as any other of the frequently used procedures (10) or (11). Furthermore, it enables us to control the whole rounding process.

In the linearly ordered set $\{\hat{\mathbb{R}},\leqq\}$ let us consider now the groupoid $\{\hat{\mathbb{R}},*\}$, where $*$ denotes one of the signs $\{+,-,\times,/\}$. We know from the Theorems 3 and 4, that we can define the optimal lower resp. the optimal upper screen groupoid $\{\hat{\mathbb{R}}_M,\underset{\vee}{*}\}$ resp. $\{\hat{\mathbb{R}}_M,\hat{*}\}$ by

$$\bigwedge_{a\in\hat{\mathbb{R}}_M}\bigwedge_{b\in\hat{\mathbb{R}}_M}a\underset{\vee}{*}b:=\nabla(a*b) \quad \text{resp.} \quad \bigwedge_{a\in\hat{\mathbb{R}}_M}\bigwedge_{b\in\hat{\mathbb{R}}_M}a\hat{*}b:=\Delta(a*b). \tag{13}$$

Because of Theorem 8 every further optimal screen groupoid $\{\hat{\mathbb{R}}_M,\overline{*}\}$ can be described by the optimal lower and the optimal upper screen groupoids $\{\hat{\mathbb{R}}_M,\underset{\vee}{*}\}$ and $\{\hat{\mathbb{R}}_M,\hat{*}\}$ for every $*\in\{+,-,\times,/\}$.

We know from Theorem 4 that the operations $\underset{\vee}{*}$, $\hat{*}$ and every further optimal operation $\overline{*}$, $*\in\{+,-,\times,/\}$, satisfy the following properties, which in the cases of $\underset{\vee}{*}$ and $\hat{*}$ even define these operations:

(RG1) $\bigwedge_{a,b,c,d\in\hat{\mathbb{R}}_M}(a*b\leqq c*d\Rightarrow a\underset{\vee}{*}b\leqq c\underset{\vee}{*}d \quad \text{resp.} \quad a\hat{*}b\leqq c\hat{*}d \quad \text{resp.} \quad a\overline{*}b\leqq c\overline{*}d)$.

(RG2) $$\bigwedge_{a\in\hat{\mathbb{R}}_M}\bigwedge_{b\in\hat{\mathbb{R}}_M}(a*b\in\hat{\mathbb{R}}_M)\Rightarrow a\underset{\vee}{*}b=a\hat{*}b=a\overline{*}b=a*b).$$

(RG3) $$\bigwedge_{a\in\hat{\mathbb{R}}_M}\bigwedge_{b\in\hat{\mathbb{R}}_M}a\underset{\vee}{*}b\leqq a*b \quad \text{resp.} \quad \bigwedge_{a\in\hat{\mathbb{R}}_M}\bigwedge_{b\in\hat{\mathbb{R}}_M}a*b\leqq a\hat{*}b.$$

The operations $\underset{\vee}{*}$ and $\hat{*}$ in (13) are not independent of each other. Because of (S3) we can for instance express the subtraction by the addition:

$$a \underset{\vee}{-} b = \nabla(a-b) = \nabla\big(a+(-b)\big) = a \underset{\vee}{+} (-b) \quad \text{resp.} \quad a \hat{-} b = a \hat{+} (-b). \tag{14}$$

Because of (12) we can further express every operation $\hat{*}$, $*\in\{+, -, \times, /\}$ by the corresponding operation $\underset{\vee}{*}$ and vice versa. We get for instance

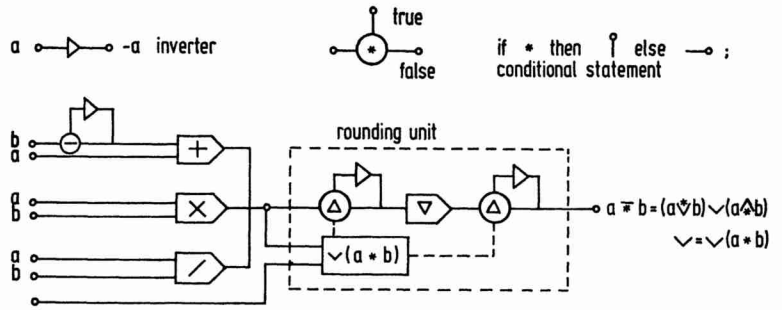
$$a \hat{*} b = \Delta(a*b) = -\nabla\big(-(a*b)\big) = \begin{cases} -\big((-a) \underset{\vee}{*} (-b)\big), & \text{for } *\in\{+, -\} \\ -\big((-a) \underset{\vee}{*} b\big), & \text{for } *\in\{\times, /\}. \end{cases} \tag{15}$$

From (14), (15) and Theorem 8 we get the important result, that an implementation for instance of the operations $\underset{\vee}{-}$, $\underset{\vee}{\times}$ and $\underset{\vee}{/}$ on a computer is already sufficient to carry out the operations of every optimal screen groupoid

$$\{\hat{\mathbb{R}}_M, \bar{*}\}, \quad *\in\{+, -, \times, /\}$$

on that computer.

Fig. 4 shows in principle the construction of the arithmetic unit of a computer which enables us to get the results of every, in the above sense, optimal screen groupoid. The "rounding unit" is the essential part of this arithmetic unit. In Fig. 4 we assumed, that the operations $+$, $\times$ and $/$ for elements of $\hat{\mathbb{R}}_M$ are available on the computer.



$a \underset{\vee}{-} b = \nabla(a + (-b))$ $\qquad$ $a \hat{-} b = -\nabla(-(a + (-b)))$

$a \underset{\vee}{+} b = \nabla(a + b)$ $\qquad$ $a \hat{+} b = -\nabla(-(a + b))$

$a \underset{\vee}{\times} b = \nabla(a \times b)$ $\qquad$ $a \hat{\times} b = -\nabla(-(a \times b))$

$a \underset{\vee}{/} b = \nabla(a/b)$ $\qquad$ $a \hat{/} b = -\nabla(-(a/b))$

Fig. 4. Arithmetic unit of an universal computer

3.2. Approximation of the Real Interval Arithmetic on an Upper Screen

Let now $[a_1; a_2]$ be a closed real interval and let us consider the sets

$$I(\hat{\mathbb{R}}) := \{[a_1; a_2] \,|\, a_1, a_2 \in \hat{\mathbb{R}}\} \quad \text{and} \quad I(\hat{\mathbb{R}}_M) := \{[a_1; a_2] \,|\, a_1, a_2 \in \hat{\mathbb{R}}_M\}.$$

Every interval computation intends to determine a set which includes the solution of a given problem. If we want to approximate the interval arithmetic in $I(\hat{\mathbb{R}})$ on a computer in a suitable way, we have to ask for two conditions:

a) the result of any computation on the computer has to include the result of the corresponding computation in $I(\hat{\mathbb{R}})$;

b) the difference between the results of the computation on the computer and that in $I(\hat{\mathbb{R}})$ has to be minimal (for the given computer).

That means, that we have to apply optimal roundings, which enlarge each interval.

Using the order relation $\subseteq$, $\{I(\hat{\mathbb{R}}), \subseteq\}$ is an ordered set and $I(\hat{\mathbb{R}}_M)$ is an upper screen of $\{I(\hat{\mathbb{R}}), \subseteq\}$.

We know from Theorem 1, that we can define the optimal, "outward" directed rounding $\diamond : I(\hat{\mathbb{R}}) \to I(\hat{\mathbb{R}}_M)$ by

$$\bigwedge_{A \in I(\hat{\mathbb{R}})} \diamond A := o\left(U_{I(\hat{\mathbb{R}}_M)}(A)\right).$$

Since the set $\hat{\mathbb{R}}$ is lineary ordered, we get immediately with $A = [a_1; a_2]$

$$\diamond A = \diamond [a_1; a_2] = [\nabla a_1; \Delta a_2]. \tag{16}$$

Because of Definition 3 the rounding $\diamond$ satisfies the following properties, which also define it:

(R1) $$\bigwedge_{A \in I(\hat{\mathbb{R}})} \bigwedge_{B \in I(\hat{\mathbb{R}})} (A \subseteq B \Rightarrow \diamond A \subseteq \diamond B).$$

(R2) $$\bigwedge_{A \in I(\hat{\mathbb{R}}_M)} \diamond A = A.$$

(R3) $$\bigwedge_{A \in I(\hat{\mathbb{R}})} A \subseteq \diamond A.$$

In $I(\hat{\mathbb{R}})$ there exist intervals which do not include intervals of $I(\hat{\mathbb{R}}_M)$, i.e. $I(\hat{\mathbb{R}}_M)$ is not a lower screen of $I(\hat{\mathbb{R}})$. Therefore we will not define a "inward"-directed rounding of $I(\hat{\mathbb{R}})$ into $I(\hat{\mathbb{R}}_M)$.

Special intervals of $I(\hat{\mathbb{R}})$ are the elements $A := [a; a]$. If $a \in \hat{\mathbb{R}}_M$, we have $\diamond [a; a] = [\nabla a; \Delta a] = [a; a]$. Yet in general $(a \in \hat{\mathbb{R}})$ we have $[a; a] \subseteq \diamond [a; a] = [\nabla a; \Delta a]$.

For elements $A, B \in \mathbb{P}(\hat{\mathbb{R}})$ the operations $* \in \{+, -, \times, /\}$ are defined by:

$$A * B := \{a * b \mid a, b \in \hat{\mathbb{R}} \wedge a \in A \wedge b \in B\}. \tag{17}$$

If $A = [a_1; a_2]$, $B = [b_1; b_2] \in I(\hat{\mathbb{R}})$ from (17) we get the following explicit rules:

$$A * B = [a_1; a_2] * [b_1; b_2] = \left[\min_{i,j=1,2}(a_i * b_j); \max_{i,j=1,2}(a_i * b_j)\right]. \tag{18}$$

In practice to get each bound of the resulting interval it is only necessary to calculate one of the four operations (18). Only in the case of multiplying two intervals, each of which includes the zero element, two of the four products have to be compared.

As a consequence of (17) we get for all $* \in \{+, -, \times, /\}$:

(G3′) $$\bigwedge_{A, B, C, D \in I(\hat{\mathbb{R}})} (A \subseteq C \wedge B \subseteq D \Rightarrow A * B \subseteq C * D),$$

i.e. for all $*\in\{+, -, \times, /\}$, $\{I(\hat{\mathbb{R}}), *, \subseteq\}$ is an ordered groupoid and (see Example 7) it has the right neutral element $[0; 0]$ in the cases of addition and subtraction and $[1; 1]$ in the cases of multiplication and division. We know from Theorems 3 and 4, that for all $*\in\{+, -, \times, /\}$ we can define the optimal upper screen groupoid $\{I(\hat{\mathbb{R}}_M), \hat{\underset{\vee}{*}}, \subseteq\}$ of $\{I(\hat{\mathbb{R}}), *, \subseteq\}$ by

$$\bigwedge_{A\in I(\hat{\mathbb{R}}_M)} \bigwedge_{B\in I(\hat{\mathbb{R}}_M)} A\hat{\underset{\vee}{*}}B := \diamond(A*B). \tag{19}$$

Because of Definition 5 the operations $\hat{\underset{\vee}{*}}$, $*\in\{+, -, \times, /\}$, satisfy the following properties, which also define them:

(RG1) $$\bigwedge_{A,B,C,D\in I(\hat{\mathbb{R}}_M)} (A*B\subseteq C*D \Rightarrow A\hat{\underset{\vee}{*}}B \subseteq C\hat{\underset{\vee}{*}}D).$$

(RG2) $$\bigwedge_{A\in I(\hat{\mathbb{R}}_M)} \bigwedge_{B\in I(\hat{\mathbb{R}}_M)} \big((A*B\in(I\hat{\mathbb{R}}_M)) \Rightarrow A\hat{\underset{\vee}{*}}B = A*B\big).$$

(RG3) $$\bigwedge_{A\in I(\hat{\mathbb{R}}_M)} \bigwedge_{B\in I(\hat{\mathbb{R}}_M)} A*B\subseteq A\hat{\underset{\vee}{*}}B.$$

As a consequence of (16), (18) and (19) we get immediately

$$A\hat{\underset{\vee}{*}}B = [a_1; a_2]\hat{\underset{\vee}{*}}[b_1; b_2] = \big[\nabla \min_{i,j=1,2}(a_i*b_j); \Delta \max_{i,j=1,2}(a_i*b_j)\big].$$

Since the operators ∇ and Δ are monotone functions, this is equivalent to

$$A\hat{\underset{\vee}{*}}B = \diamond(A*B) = [a_1; a_2]\hat{\underset{\vee}{*}}[b_1; b_2] = \big[\min_{i,j=1,2}(a_i \underset{\vee}{*} b_j); \max_{i,j=1,2}(a_i \hat{*} b_j)\big]. \tag{20}$$

In practice, to get each bound of the resulting interval, it is again only necessary to calculate one of the four products in (20). Only in the case of multiplication of two intervals, each of which includes the zero element, two of the four products have to be compared.

The interval calculus in $I(\hat{\mathbb{R}}_M)$, which is now defined, fulfills the conditions a) and b), which we stated at the beginning of this chapter.

It is a consequence of (20) that the operations $\hat{\underset{\vee}{*}}\in\{\hat{\underset{\vee}{+}}, \hat{\underset{\vee}{-}}, \hat{\underset{\vee}{\times}}, \hat{\underset{\vee}{/}}\}$ in $I(\hat{\mathbb{R}}_M)$ are available, if the operations $\underset{\vee}{*}$ and $\hat{*}$ for elements of $\hat{\mathbb{R}}_M$ are available on the computer. We have already seen, that for this the existence of the three operations $\underset{\vee}{+}$, $\underset{\vee}{\times}$, $\underset{\vee}{/}$ or an equivalent triple is sufficient. That means, that the computer in Fig. 4 also can be used for interval computations in an optimal way.

Further Properties of the Operations $\hat{\underset{\vee}{*}}$

The operators $\diamond$ and $\hat{\underset{\vee}{*}}\in\{\hat{\underset{\vee}{+}}, \hat{\underset{\vee}{-}}, \hat{\underset{\vee}{\times}}, \hat{\underset{\vee}{/}}\}$ especially fullill the properties proved in the Theorems 1 to 6. Particularly for all, $A, B\in I(\hat{\mathbb{R}})$ the inequality $\diamond(A*B)\subseteq(\diamond A)\hat{\underset{\vee}{*}}(\diamond B)$ holds, which means, that the mapping $\diamond$ is not homomorphic. We show now with a simple example, that the inclusion sign really occurs.

Without loss of generality we may assume a computer with the number representation $m\cdot 10^e$ and the range of numbers $m = -9(1)9$ and $e = -1(1)1$.

Let be $A := [3.6; 5.4]$ and $B := [3.5; 4.5] \in I(\hat{\mathbb{R}})$. Then we get $A + B = [7.1; 9.9]$ and with (16) $\diamond A = [3; 6]$, $\diamond B = [3; 5]$ and $\diamond (A + B) = [7; 1 \cdot 10]$. Yet $(\diamond A) \hat{\dagger}_{\vee} (\diamond B) = \diamond \big((\diamond A) + (\diamond B)\big) = \diamond [6; 11] = [6; 2 \cdot 10]$ and therefore $\diamond (A + B) = [7; 1 \cdot 10] \subset [6; 2 \cdot 10] = (\diamond A) \hat{*}_{\vee} (\diamond B)$.

From definition (17) of the arithmetic in $\mathbb{P}(\hat{\mathbb{R}})$ it follows easily, that the addition and the multiplication in $\mathbb{P}(\hat{\mathbb{R}})$ are associative. We show now with simple examples, that these associative laws in $I(\hat{\mathbb{R}})$ are lost in the arithmetic in $I(\hat{\mathbb{R}}_M)$. We take again the special number representation, used in the above example. Let be $A, B, C \in I(\hat{\mathbb{R}}_M)$.

a) Let be $A := [3; 6]$, $B := [3; 5]$, $C := [3; 4]$. Then we get

$$A \hat{\dagger}_{\vee} (B \hat{\dagger}_{\vee} C) = [3; 6] \hat{\dagger}_{\vee} [6; 9] = \diamond [9; 15] = \diamond [9; 2 \cdot 10].$$

$$(A \hat{\dagger}_{\vee} B) \hat{\dagger}_{\vee} C = \diamond [6; 11] \hat{\dagger}_{\vee} [3; 4] = [6; 2 \cdot 10] \hat{\dagger}_{\vee} [3; 4] = \diamond [9; 24] = [9; 3 \cdot 10]$$

and therefore $A \hat{\dagger}_{\vee} (B \hat{\dagger}_{\vee} C) \neq (A \hat{\dagger}_{\vee} B) \hat{\dagger}_{\vee} C$.

b) Let be $A := [3; 4]$, $B := [4; 5]$, $C := [4; 4]$. Then we get

$$A \hat{\times}_{\vee} (B \hat{\times}_{\vee} C) = [3; 4] \hat{\times}_{\vee} (\diamond [16; 20]) = [3; 4] \hat{\times}_{\vee} [1 \cdot 10; 2 \cdot 10] = [3 \cdot 10; 8 \cdot 10].$$

$$(A \hat{\times}_{\vee} B) \hat{\times}_{\vee} C = \diamond [12; 20] \hat{\times}_{\vee} [4; 4] = [1 \cdot 10; 2 \cdot 10] \hat{\times}_{\vee} [4; 4] = [4 \cdot 10; 8 \cdot 10],$$

and therefore $A \hat{\times}_{\vee} (B \hat{\times}_{\vee} C) \neq (A \hat{\times}_{\vee} B) \hat{\times}_{\vee} C$.

References

1. Apostolatos, N., Christ, H., Santo, H., Wipperman, H.: Rounding control and the algorithmic language ALGOL-68. Report, University of Karlsruhe, Computer Center, July 1968, S. 1–9.
2. — Kulisch, U.: Grundlagen einer Maschinenintervallarithmetik. Computung **2**, Fasc. 2, 89–104 (1967).
3. Christ, H.: Realisierung einer Maschinenintervallarithmetik auf beliebigen ALGOL-60 Compilern. Elektronische Rechenanlagen **10**, H. 5, 217–222 (1968).
4. Herzberger, J.: Metrische Eigenschaften von Mengensystemen und einige Anwenwendungen. Dr.-Dissertation, University of Kalsruhe, 1969, S. 1–49.
5. Kulisch, U.: Grundzüge der Intervallrechnung, Überblicke Mathematik. Bibliographisches Institut, Mannheim 1969, S. 51–98.
6. Rutishauser, H.: Versuch einer Axiomatik des Numerischen Rechnens. Kurzvortrag, GAMM-Tagung, Aachen 1969.

Prof. Dr. U. Kulisch
Mathematics Research Center
The University of Wisconsin
Madison, Wisconsin 53706
U.S.A.
and
Institut für Angewandte Mathematik
Universität Karlsruhe
BRD-7500 Karlsruhe
Englerstraße 2
Deutschland