On realization of floating-point directed interval arithmetic

Andrzej Marciniak

Poznań University of Technology, Institute of Computing Science, Piotrowo 2, 60-965 Poznań, Poland

e-mail: Andrzej.Marciniak@put.poznan.pl

(September 6, 2012)

Abstract : In the paper we present a realization of basic arithmetic operations in floating-point directed interval arithmetic which differs from the one presented e.g. in [9] and which is included in the PASCAL–XSC programming language. Our approach guarantees obtaining only one resulting interval in each floating-point operation regardless of intervals (proper or directed) used as operands. This interval includes all possible rounding errors.

Key words: interval arithmetic, directed interval arithmetic, floating-point directed interval arithmetic, estimation of rounding errors

I. INTRODUCTION

Floating-point interval arithmetic is a way for automatic estimation of rounding errors. Using machine intervals we can also represent any real number (in mathematical sense) which is not exactly represented in computers in the form of interval. It is sufficient to represent such a number as the interval with the ends being two succeeding machine numbers between the given real number is included.

There are known two interval arithmetic. The first one operates only on proper intervals, i.e. on intervals with left ends less or equal to the right ends. In th second interval arithmetic, called directed interval arithmetic, an interval is a set of ordered couples of finite real numbers and the left end of interval can be greater than the right one. Of course, a realization of directed interval arithmetic on computers is more complicated than proper interval arithmetic.

The main reasons for developing directed interval arithmetic are such that in proper interval arithmetic there are not exist inverse elements with respect to the addition and to the multiplication of intervals. It means that the solutions of equations

$$A + X = B$$

and

$$A \times X = B,$$

where A and B are known intervals and X is unknown, can not be solved in general. But in a number of problems such equations or system of equations arise and should be solved.

An realization of proper interval arithmetic on computers is documented, among others, in [2] – [6]. In [4] and [5] one can find a description of implementations of this arithmetic in the C–XSC and PASCAL–XSC computer languages, respectively. Details concerning directed interval arithmetic and its applications can be found in [7], while its implementation in IEEE floating-point environment has been presented in [9].

The implementation presented in [9] consists in finding for any single arithmetic operation two intervals: one with so called outward rounding and one with inward rounding, and there is a problem with choosing only one interval for further calculations. Since theoretically we can consider also two other resulting intervals, we propose to chose an appropriate resulting interval on the basis of the widths of intervals and to provide all calculations using *Extended* real type (accessible e.g. in the Delphi programming language) instead IEEE *Double* real type. For this purpose we have extended our Delphi *IntervalArithmetic* unit, given in [6], to perform all basic arithmetic operations on improper intervals.

In the paper we present our approach for floating-point basic arithmetic operations on directed intervals in details using a pseudocode.

II. BASIC ARITHMETIC OPERATIONS

Let the set H defined by

$$H = \{[a,b]:\ a,b \in R\} = IR \cup \overline{IR},\ \text{ where } \overline{IR} = \{[a^-, a^+]:\ a^- \geq a^+;\ a^-, a^+ \in R\}$$

be the set of all ordered couples of finite real numbers. Moreover, let us denote

$$\mathrm{T} = \{A \in IR:\ a^- a^+ \leq 0\} \cup \{A \in \overline{IR}:\ a^- a^+ \leq 0\} = Z \cup \overline{Z}.$$

For a directed interval $A = [a^-, a^+]$ let us define the sign operator σ by

$$\sigma(A) = \begin{cases} +, & \text{if } 0 \leq a^- \text{ and } 0 \leq a^+, \\ -, & \text{if } a^- \leq 0 \text{ and } a^+ \leq 0, \text{ but } A \neq [0, 0], \end{cases}$$

and a binary variable direction operator τ by

$$\tau(A) = \begin{cases} +, & \text{if } a^- \leq a^+, \\ -, & \text{otherwise.} \end{cases}$$

According to [7] and [9], the addition in H is defined as follows:

$$A + B = [a^- + b^-, a^+ + b^+],\ \text{ for }\ A, B \in H,$$

For multiplication in H we have

$$A \times B = \begin{cases} [a^{-\sigma(B)} b^{-\sigma(A)}, a^{\sigma(B)\sigma(A)}], \text{ for } A, B \in H \setminus \mathrm{T}, \\ [a^{\sigma(A)\tau(B)} b^{-\sigma(A)}, a^{\sigma(A)\tau(B)} b^{\sigma(A)}], \text{ for } A \in H \setminus \mathrm{T}, B \in \mathrm{T}, \\ [a^{-\sigma(B)} b^{\sigma(B)\tau(A)}, a^{\sigma(B)} b^{\sigma(B)\tau(A)}], \text{ for } A \in \mathrm{T}, B \in H \setminus \mathrm{T}, \\ [\min\{a^- b^+, a^+ b^-\}, \max\{a^- b^-, a^+ b^+\}], \text{ for } A, B \in Z, \\ [\max\{a^- b^-, a^+ b^+\}, \min\{a^- b^+, a^+ b^-\}], \text{ for } A, B \in \overline{Z}, \\ 0, \text{ for } A \in Z, B \in \overline{Z} \text{ or } A \in \overline{Z}, B \in Z \end{cases}.$$

From the definition of multiplication for $B \in H$ we obtain

$$(-1) \times B = [-b^+, -b^-] = -B.$$

Thus, the subtraction can be defined as

$$A - B = A + (-B) = [a^- - b^+, a^+ - b^-], \ A, B \in H.$$

Let us note that in H there exist inverse elements with respect to the operations + and ×. Namely, we have

$$-_h A = [-a^-, -a^+], \text{ for } A \in H,$$
$$1/_h A = [1/a^-, 1/a^+], \text{ for } A \in H \setminus \mathrm{T}.$$

Moreover, for $A = [a^-, a^+] \in H \setminus \mathrm{T}$ there also exists a set inversion operator

$$1/A = 1/_h A_- = [1/a^+, 1/a^-],$$

where $A_- = [a^+, a^-]$, such that $1/_h (1/A) = 1/(1/_h A) = A_-$. Thus, we can define the division as follows:

$$A/B = A \times (1/B) = \begin{cases} [a^{-\sigma(B)} / b^{\sigma(A)}, a^{\sigma(B)} / b^{-\sigma(A)}], \text{ for } A, B \in H \setminus \mathrm{T}, \\ [a^{-\sigma(B)} / b^{-\sigma(B)\tau(A)}, a^{\sigma(B)} / b^{-\sigma(B)\tau(A)}, \text{ for } A \in \mathrm{T}, B \in H \setminus \mathrm{T}. \end{cases}$$

It should be mention that to every directed interval $A = [a^-, a^+] \in H$ we can assign a proper interval pro(*A*) with (see [7])

$$\mathrm{pro}(A) = \begin{cases} [a^-, a^+], \text{ if } \tau(A) = +, \\ [a^+, a^-], \text{ if } \tau(A) = -. \end{cases}$$

Such an interval is called the projection of *A* on a the set of proper intervals or the proper projection of *A*.

In a number of papers, among others in [7] and [9], one can find a lot of interesting properties of directed intervals. We omit a discussion of them because they are out of scope of the topic of this paper.

III. FLOATING-POINT ARITHMETIC OPERATIONS

A realization of directed interval arithmetic on computers has been presented in [9]. Omitting some details, this realization consists in finding two resulting intervals for any operation $\circ \in \{+, -, \times, /\}$ and machine intervals A and B[1]:

$$\Diamond(A \circ B) = [\nabla(A \circ B)^{-}, \Delta(A \circ B)^{+}],$$
$$\mathrm{O}(A \circ B) = [\Delta(A \circ B)^{-}, \nabla(A \circ B)^{+}].$$

The operator $\Diamond$ represents so called outward rounding and the operator O – inward rounding. The symbol ∇ is used for rounding toward $-\infty$ or downwardly directed and the symbol Δ – for rounding toward $+\infty$ or upwardly directed. An implementation of such defined interval arithmetic operations in the PASCAL–XSC language supporting IEEE floating-point standard is described in details in [9].

From the above formulas it follows that performing any single arithmetic operation we obtain two intervals and there is a problem with choosing only one interval for further calculations. Moreover, from the theoretical background one can consider two other resulting intervals, namely

$$[\nabla(A \circ B)^{-}, \nabla(A \circ B)^{+}] \text{ and } [\Delta(A \circ B)^{-}, \Delta(A \circ B)^{+}].$$

Our proposition consists in choosing an appropriate resulting interval on the basis of the widths of intervals and in providing all calculations using *Extended* real type[2]. From all possible resulting intervals we always choose the worst case, i.e. the interval with the largest width. Below we present our approach for floating-point basic arithmetic operations on directed intervals in details using a pseudocode.

First, let us introduce the floating-point **width** w of an interval $A = [a^{-}, a^{+}]$:

```
$w := \Delta(a^{+} - a^{-})$
**if** $w < 0$
  **then** $w := -w$
$\overline{w} := \nabla(a^{+} - a^{-})$
**if** $\overline{w} < 0$
  **then** $\overline{w} := -\overline{w}$
**if** $w < \overline{w}$
  **then** $w := \overline{w}$
```

A realization of **addition** for $A = [a^{-}, a^{+}]$ and $B = [b^{-}, b^{+}]$ may look as follows:

```
**if** $a^{-} \le a^{+}$ **and** $b^{-} \le b^{+}$ (proper intervals)
  **then** $A + B = [\nabla(a^{-} + b^{-}), \Delta(a^{+} + b^{+})]$
  **else begin**
        $c^{-} := \nabla(a^{-} + b^{-})$, $c^{+} := \Delta(a^{+} + b^{+})$
```

[1] A machine interval is an interval which both ends are exactly represented in a computer.

[2] The *Extended* type of real numbers are available e.g. in Delphi Pascal. This type has larger precision and range than *Double* type in IEEE standards.

```
        $d^- := \Delta(a^- + b^-)$,  $d^+ := \nabla(a^+ + b^+)$
        calculate the width $w_1$ of $[c^-, c^+]$
        calculate the width $w_2$ of $[d^-, d^+]$
        if $w_1 \ge w_2$
         then $A + B := [c^-, c^+]$
         else $A + B := [d^-, d^+]$
       end
```

For the **subtraction** we have:

```
if $a^- \le a^+$ and $b^- \le b^+$ (proper intervals)
 then $A - B := [\nabla(a^- - b^+), \Delta(a^+ - b^-)]$
 else begin
        $c^- := \nabla(a^- - b^+)$,  $c^+ := \Delta(a^+ - b^-)$
        $d^- := \Delta(a^- - b^+)$,  $d^+ := \nabla(a^+ - b^-)$
        calculate the width $w_1$ of $[c^-, c^+]$
        calculate the width $w_2$ of $[d^-, d^+]$
        if $w_1 \ge w_2$
         then $A - B := [c^-, c^+]$
         else $A - B := [d^-, d^+]$
       end
```

The **multiplication** is more complicated. We have:

```
if $a^- \le a^+$ and $b^- \le b^+$ (proper intervals)
 then $A \times B = [\min\{\nabla(a^-b^-), \nabla(a^-b^+), \nabla(a^+b^-), \nabla(a^+b^+)\},$
                $\max\{\Delta(a^-b^-), \Delta(a^-b^+), \Delta(a^+b^-), \Delta(a^+b^+)\}]$
 else if ($a^- < 0$ and $a^+ < 0$ or $a^- > 0$ and $a^+ > 0$)
      and ($b^- < 0$ and $b^+ < 0$ or $b^- > 0$ and $b^+ > 0$)
        then if $a^- > 0$ and $a^+ > 0$ and $b^- > 0$ and $b^+ > 0$
              then begin
                     $c^- := \nabla(a^-b^-)$,  $c^+ := \Delta(a^+b^+)$
                     $d^- := \Delta(a^-b^-)$,  $d^+ := \nabla(a^+b^+)$
                     calculate $A \times B$
                   end
             else if $a^- > 0$ and $a^+ > 0$ and $b^- < 0$ and $b^+ < 0$
                   then begin
                          $c^- := \nabla(a^+b^-)$,  $c^+ := \Delta(a^-b^+)$
                          $d^- := \Delta(a^+b^-)$,  $d^+ := \nabla(a^-b^+)$
                          calculate $A \times B$
                        end
                  else if $a^- < 0$ and $a^+ < 0$ and $b^- > 0$ and $b^+ > 0$
                        then begin
                               $c^- := \nabla(a^-b^+)$,  $c^+ := \Delta(a^+b^-)$
                               $d^- := \Delta(a^-b^+)$,  $d^+ := \nabla(a^+b^-)$
```

```
                    calculate $A \times B$
                   end
              else begin
                    $c^- := \nabla(a^+b^+)$, $c^+ := \Delta(a^-b^-)$
                    $d^- := \Delta(a^+b^+)$, $d^+ := \nabla(a^-b^-)$
                    calculate $A \times B$
                  end
else if $(a^- < 0$ and $a^+ < 0$ or $a^- > 0$ and $a^+ > 0)$
       and $(b^- \le 0$ and $b^+ \ge 0$ or $b^- \ge 0$ and $b^+ \le 0)$
    then if $a^- > 0$ and $a^+ > 0$ and $b^- \le b^+$
          then begin
               $c^- := \nabla(a^+b^-)$, $c^+ := \Delta(a^+b^+)$
               $d^- := \Delta(a^+b^-)$, $d^+ := \nabla(a^+b^+)$
               calculate $A \times B$
              end
          else if $a^- > 0$ and $a^+ > 0$ and $b^- > b^+$
              then begin
                    $c^- := \nabla(a^-b^-)$, $c^+ := \Delta(a^-b^+)$
                    $d^- := \Delta(a^-b^-)$, $d^+ := \nabla(a^-b^+)$
                    calculate $A \times B$
                   end
              else if $a^- < 0$ and $a^+ < 0$ and $b^- \le b^+$
                   then begin
                         $c^- := \nabla(a^-b^+)$, $c^+ := \Delta(a^-b^-)$
                         $d^- := \Delta(a^-b^+)$, $d^+ := \nabla(a^-b^-)$
                         calculate $A \times B$
                        end
                   else begin
                         $c^- := \nabla(a^+b^+)$, $c^+ := \Delta(a^+b^-)$
                         $d^- := \Delta(a^+b^+)$, $d^+ := \nabla(a^+b^-)$
                         calculate $A \times B$
                        end
    else if $(a^- \le 0$ and $a^+ \ge 0$ or $a^- \ge 0$ and $a^+ \le 0)$
           and $(b^- < 0$ and $b^+ < 0$ or $b^- > 0$ and $b^+ > 0)$
         then if $a^- \le a^+$ and $b^- > 0$ and $b^+ > 0$
              then begin
                    $c^- := \nabla(a^-b^+)$, $c^+ := \Delta(a^+b^+)$
                    $d^- := \Delta(a^-b^+)$, $d^+ := \nabla(a^+b^+)$
                    calculate $A \times B$
                   end
              else if $a^- \le a^+$ and $b^- < 0$ and $b^+ < 0$
                   then begin
                         $c^- := \nabla(a^+b^-)$, $c^+ := \Delta(a^-b^-)$
```

```
                $d^- := \Delta(a^+b^-)$, $d^+ := \nabla(a^-b^-)$
                calculate $A \times B$
              end
          else if $a^- > a^+$ and $b^- > 0$ and $b^+ > 0$
              then begin
                    $c^- := \nabla(a^-b^-)$, $c^+ := \Delta(a^+b^-)$
                    $d^- := \Delta(a^-b^-)$, $d^+ := \nabla(a^+b^-)$
                    calculate $A \times B$
                  end
              else begin
                    $c^- := \nabla(a^+b^+)$, $c^+ := \Delta(a^-b^+)$
                    $d^- := \Delta(a^+b^+)$, $d^+ := \nabla(a^-b^+)$
                    calculate $A \times B$
                  end
else if $a^- \geq 0$ and $a^+ \leq 0$ and $b^- \geq 0$ and $b^+ \leq 0$
    then begin
          $c_1^- := \nabla(a^-b^-)$, $c_2^- := \nabla(a^+b^+)$
          if $c_1^- \leq c_2^-$
            then $c^- := c_2^-$
            else $c^- := c_1^-$
          $c_1^+ := \Delta(a^-b^+)$, $c_2^+ := \Delta(a^+b^-)$
          if $c_1^+ \leq c_2^+$
            then $c^+ := c_1^+$
            else $c^+ := c_2^+$
          $d_1^- := \Delta(a^-b^-)$, $d_2^- := \Delta(a^+b^+)$
          if $d_1^- \leq d_2^-$
            then $d^- := d_2^-$
            else $d^- := d_1^-$
          $d_1^+ := \nabla(a^-b^+)$, $d_2^+ := \nabla(a^+b^-)$
          if $d_1^+ \leq d_2^+$
            then $d^+ := d_1^+$
            else $d^+ := d_2^+$
          calculate $A \times B$
        end
    else $A \times B := [0, 0]$
```

where the expression “calculate $A \times B$” means a function which can be described as follows:

```
calculate the width $w_1$ of $[c^-, c^+]$
calculate the width $w_2$ of $[d^-, d^+]$
if $w_1 \geq w_2$
  then $A \times B := [c^-, c^+]$
  else $A \times B := [d^-, d^+]$
```

Finally, for the **division** we have:

```
if $a^- \le a^+$ and $b^- \le b^+$ (proper intervals)
 then $A/B := [\min\{\nabla(a^-/b^-), \nabla(a^-/b^+), \nabla(a^+/b^-), \nabla(a^+/b^+)\},$
            $\max\{\Delta(a^-/b^-), \Delta(a^-/b^+), \Delta(a^+/b^-), \Delta(a^+/b^+)\}]$
 else if $(a^- < 0$ and $a^+ < 0$ or $a^- > 0$ and $a^+ > 0)$
      and $(b^- < 0$ and $b^+ < 0$ or $b^- > 0$ and $b^+ > 0)$
    then if $a^- > 0$ and $a^+ > 0$ and $b^- > 0$ and $b^+ > 0$
         then begin
                $c^- := \nabla(a^-/b^+)$, $c^+ := \Delta(a^+/b^-)$
                $d^- := \Delta(a^-/b^+)$, $d^+ := \nabla(a^+/b^-)$
                calculate $A/B$
              end
         else if $a^- > 0$ and $a^+ > 0$ and $b^- < 0$ and $b^+ < 0$
              then begin
                     $c^- := \nabla(a^+/b^+)$, $c^+ := \Delta(a^-/b^-)$
                     $d^- := \Delta(a^+/b^+)$, $d^+ := \nabla(a^-/b^-)$
                     calculate $A/B$
                   end
              else if $a^- < 0$ and $a^+ < 0$ and $b^- > 0$ and $b^+ > 0$
                   then begin
                          $c^- := \nabla(a^-/b^-)$, $c^+ := \Delta(a^+/b^+)$
                          $d^- := \Delta(a^-/b^-)$, $d^+ := \nabla(a^+/b^+)$
                          calculate $A/B$
                        end
                   else begin
                          $c^- := \nabla(a^+/b^-)$, $c^+ := \Delta(a^-/b^+)$
                          $d^- := \Delta(a^+/b^-)$, $d^+ := \nabla(a^-/b^+)$
                          calculate $A/B$
                        end
    else if $(a^- \le 0$ and $a^+ \ge 0$ or $a^- \ge 0$ and $a^+ \le 0)$
          and $(b^- < 0$ and $b^+ < 0$ or $b^- > 0$ and $b^+ > 0)$
       then if $a^- \le a^+$ and $b^- > 0$ and $b^+ > 0$
            then begin
                   $c^- := \nabla(a^-/b^-)$, $c^+ := \Delta(a^+/b^-)$
                   $d^- := \Delta(a^-/b^-)$, $d^+ := \nabla(a^+/b^-)$
                   calculate $A/B$
                 end
            else if $a^- \le a^+$ and $b^- < 0$ and $b^+ < 0$
                 then begin
                        $c^- := \nabla(a^+/b^+)$, $c^+ := \Delta(a^-/b^+)$
                        $d^- := \Delta(a^+/b^+)$, $d^+ := \nabla(a^-/b^+)$
                        calculate $A/B$
                      end
                 else if $a^- > a^+$ and $b^- > 0$ and $b^+ > 0$
```

```
                    then begin
                            $c^- := \nabla(a^- / b^+)$,  $c^+ := \Delta(a^+ / b^+)$
                            $d^- := \Delta(a^- / b^+)$,  $d^+ := \nabla(a^+ / b^+)$
                            calculate  $A / B$
                        end
                    else begin
                            $c^- := \nabla(a^+ / b^-)$,  $c^+ := \Delta(a^- / b^-)$
                            $d^- := \Delta(a^+ / b^-)$,  $d^+ := \nabla(a^- / b^-)$
                            calculate  $A / B$
                        end
            else error “division by interval containing zero”
```

where “calculate A / B ” stands for a function described as follows:

```
calculate the width  $w_1$  of  $[c^-, c^+]$
calculate the width  $w_2$  of  $[d^-, d^+]$
if  $w_1 \geq w_2$
  then  $A / B := [c^-, c^+]$
  else  $A / B := [d^-, d^+]$
```

All the above operations have been implemented in our *IntervalArithmetic* unit written in the Delphi Pascal programming language. This unit, still developed, can be loaded from [1].

References

[1] Delphi Pascal *IntervalArithmetic* unit, http://www.cs.put.poznan.pl/amarciniak/DEL-wyklady/IntervalArithmetic.pas.

[2] Hammer, R., Hocks, M., Kulisch, U., Ratz, D., *Numerical Toolbox for Verified Computing I: Basic Numerical Problems*, Springer, Berlin 1993.

[3] Jaulin, L., Kieffer, M., Didrit, O., Walter, É., *Applied Interval Analysis*, Springer-Verlag, London 2001.

[4] Klatte, R., Kulisch, U., Lawo, C., Rauch, M., Wiethoff, A., *C–XSC. A C++ Class Library for Extended Scientific Computing*, Springer-Verlag, Berlin 1993.

[5] Klatte, R., Kulisch, U., Neaga, M., Ratz, D., Ullrich, Ch., *PASCAL–XSC. Language Reference with Examples*, Springer-Verlag, Berlin 1992.

[6] Marciniak, A., *Selected Interval Methods for Solving the Initial Value Problem*, Publishing House of Poznan University of Technology, Poznan 2009.

[7] Markov, S., On Directed Interval Arithmetic and its Applications, *Journal of Universal Computer Science* 7 (1995), 514–526.

[8] Moore, R. E., Kearfott, R. B., Cloud, M. J., *Introduction to Interval Analysis*, SIAM, Philadelphia 2009.

[9] Popova, E. D., Extended Interval Arithmetic in IEEE Floating-Point Environment, *Interval Computations* 4 (1994), 100–129.

[10] Shokin, J. I., *Interval Analysis* [in Russian], Nauka, Novosibirsk 1982.