Development of a library for a symbolic floating-point arithmetic

C.-P. Jeannerod, N. Louvet, J.-M. Muller, **Antoine Plet**

LIP at ENS de Lyon

28th June, 2016

RAIM 2016, Banyuls-sur-mer

Outline

Floating-point numbers (base β, precision p)

$$x = (-1)^s \cdot m \cdot \beta^{e-p+1}$$

- $s \in \{0, 1\}$,
- $m \in \mathbb{N}$ with $1 \leqslant m < \beta^p$
- $e \in \mathbb{Z}$ with $e_{min} \leqslant e \leqslant e_{max}$

Floating-point numbers (base β, precision p)

$$x = (-1)^s \cdot m \cdot \beta^{e-p+1}$$

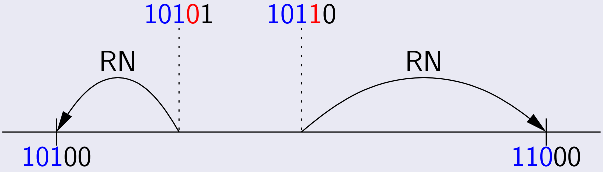
- $s \in \{0, 1\}$,
- $m \in \mathbb{N}$ (significand) with $\beta^{p-1} \leqslant m < \beta^p$
- $e \in \mathbb{Z}$ (exponent)

Floating-point numbers (base β, precision p)

$$x = (-1)^s \cdot m \cdot \beta^{e-p+1}$$

- $s \in \{0, 1\}$,
- $m \in \mathbb{N}$ (significand) with $\beta^{p-1} \leqslant m < \beta^p$
- $e \in \mathbb{Z}$ (exponent)

Rounding to nearest (tiesToEven, $\beta = 2$, $p = 3$)

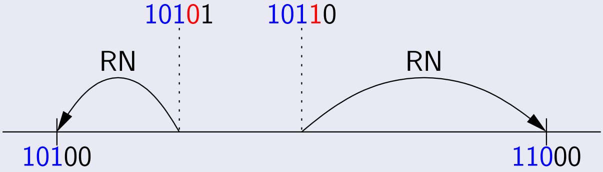


Floating-point numbers (base β, precision p)

$$x = (-1)^s \cdot m \cdot \beta^{e-p+1}$$

- $s \in \{0,1\}$,
- $m \in \mathbb{N}$ (significand) with $\beta^{p-1} \leqslant m < \beta^p$
- $e \in \mathbb{Z}$ (exponent)

Rounding to nearest (tiesToEven, $\beta = 2$, $p = 3$)



$+$, $-$, $\times$, $\div$ and fma (fused multiply-add, computes $ab + c$).
Correct rounding: compute the rounding of the exact result.

Example: Kahan's algorithm for evaluating $x = ad - bc$

algorithm Kahan(a, b, c, d)
$\widehat{w} \leftarrow \mathrm{RN}(bc)$;
$e \leftarrow \mathrm{RN}(\widehat{w} - bc)$;
$\widehat{f} \leftarrow \mathrm{RN}(ad - \widehat{w})$;
$\widehat{x} \leftarrow \mathrm{RN}(\widehat{f} + e)$;

- a, b, c, d are floating-point numbers
- $\widehat{x}$ is the computed result
- $u = \frac{1}{2}\beta^{1-p}$ is the unit roundoff

$\hookrightarrow$ Relative error bound [JLM13a]: $\frac{|\widehat{x}-x|}{|x|} \leqslant 2u$

Example: Kahan's algorithm for evaluating $x = ad - bc$

algorithm Kahan(a, b, c, d)
$\widehat{w} \leftarrow \mathrm{RN}(bc)$;
$e \leftarrow \mathrm{RN}(\widehat{w} - bc)$;
$\widehat{f} \leftarrow \mathrm{RN}(ad - \widehat{w})$;
$\widehat{x} \leftarrow \mathrm{RN}(\widehat{f} + e)$;

- a, b, c, d are floating-point numbers
- $\widehat{x}$ is the computed result
- $u = \frac{1}{2}\beta^{1-p}$ is the unit roundoff

$\hookrightarrow$ Relative error bound [JLM13a]: $\frac{|\widehat{x}-x|}{|x|} \leqslant 2u$

Asymptotic optimality of the relative error bound $2u$ [JLM13a]

Inputs parametrized by β and p:

$$a = b = \beta^{p-1} + 1$$
$$c = \beta^{p-1} + \frac{\beta}{2}\beta^{p-2}$$
$$d = 2\beta^{p-1} + \frac{\beta}{2}\beta^{p-2}$$

Relative error on the result is:

$$\frac{|\widehat{x} - x|}{|x|} = \frac{2u}{1+2u} \sim 2u \text{ as } p \to \infty.$$

$\hookrightarrow$ Symbolic floating-point numbers

Example: Kahan's algorithm for evaluating $x = ad - bc$

Paper-and-pencil calculations with symbolic floating-point numbers can be tedious: we propose to manipulate such numbers in a computer algebra system.

First step in Kahan's algorithm:

$$\begin{aligned} b &= \beta^{p-1} + 1 \\ c &= \beta^{p-1} + \tfrac{\beta}{2}\beta^{p-2} \\ bc &= \beta^{2p-2} + \tfrac{\beta}{2}\beta^{2p-3} + \beta^{p-1} + \tfrac{\beta}{2}\beta^{p-2} \\ \mathrm{RN}_p(bc) &= \beta^{2p-2} + \tfrac{\beta}{2}\beta^{2p-3} + 2\beta^{p-1} \end{aligned}$$

Example: Kahan's algorithm for evaluating $x = ad - bc$

Paper-and-pencil calculations with symbolic floating-point numbers can be tedious: we propose to manipulate such numbers in a computer algebra system.

First step in Kahan's algorithm:

$$b = \beta^{p-1} + 1$$
$$c = \beta^{p-1} + \frac{\beta}{2}\beta^{p-2}$$
$$bc = \beta^{2p-2} + \frac{\beta}{2}\beta^{2p-3} + \beta^{p-1} + \frac{\beta}{2}\beta^{p-2}$$
$$\mathrm{RN}_p(bc) = \beta^{2p-2} + \frac{\beta}{2}\beta^{2p-3} + 2\beta^{p-1}$$

$\hookrightarrow$ two sets and a rounding function:

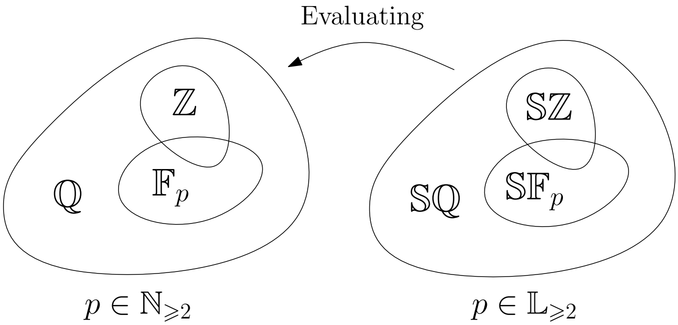
- $\mathbb{SQ}$ containing exact results of computations such as bc;
- $\mathbb{SF}_p$ containing the symbolic floating-point numbers in precision p;
- a rounding function from $\mathbb{SQ}$ to $\mathbb{SF}_p$.

Outline

β is an **even fixed** base; k is a **symbolic** variable.

We define

- $\mathbb{L} = \{ak + b \ : \ a, b \in \mathbb{Z}\}$;
- $\mathbb{E} = \{\sum_i c_i \beta^{e_i} \ : \ |c_i| \in \{1, 2, \ldots, \beta - 1\},\ e_i \in \mathbb{L}\}$;
- $\mathbb{SQ} = \mathrm{Frac}(\mathbb{E})$: **stable** by $+$, $-$, $\times$ and $\div$.



$$\mathbb{SF}_p = \{\pm m \cdot \beta^e \ : \ m \in \mathbb{SZ},\ \beta^{p-1} \leqslant |m| < \beta^p,\ e \in \mathbb{L}\}$$

Numerical rounding

$\beta = 2$, $p = 5$, and $\alpha = 85.5 = (1010101.1)_2$. $\mathrm{RN}_p(\alpha) =$?

- Unit in the last place: $\alpha = (1010101.1)_2$.

$\hookrightarrow$ $\mathrm{ulp}_p(\alpha) = 2^2$.

$\hookrightarrow$ $\mathrm{RN}_p(\alpha) = 2^2 \cdot \lfloor \alpha / \mathrm{ulp}_p(\alpha) \rceil$.

- $\alpha / \mathrm{ulp}_p(\alpha) = (10101.011)_2$.
- $\lfloor \alpha / \mathrm{ulp}_p(\alpha) \rceil = (10101)_2$.

$\hookrightarrow$ $\mathrm{RN}_p(\alpha) = (1010100)_2 = 84$.

Rounding to the nearest, in precision $p \geqslant 2$, the expression [JLM13b]:

$$f(p) = \frac{2^{3p/2} + 5 \cdot 2^{p-1}}{2^{3p} + 2^{5p/2+1}}, \quad \text{with } p \textbf{ even}.$$

$$\mathrm{RN}_2(f(2)) = 2^{-3}, \quad \mathrm{RN}_4(f(4)) = 2^{-6} + 2^{-9}, \quad \mathrm{RN}_{24}(f(24)) = 2^{-36} + 2^{-49};$$

$$\mathrm{RN}_p(f(p)) = ?$$

Rounding to the nearest, in precision $p \geqslant 2$, the expression [JLM13b]:

$$f(p) = \frac{2^{3p/2} + 5 \cdot 2^{p-1}}{2^{3p} + 2^{5p/2+1}}, \quad \text{with } p \textbf{ even}.$$

$$\mathrm{RN}_2(f(2)) = 2^{-3}, \quad \mathrm{RN}_4(f(4)) = 2^{-6} + 2^{-9}, \quad \mathrm{RN}_{24}(f(24)) = 2^{-36} + 2^{-49};$$

$$\mathrm{RN}_p(f(p)) = \ ?$$

Changes of variable:

- $p = 2k$, with $k \in \mathbb{N}^*$, gives

$$f(k) = \frac{2^{3k} + 5 \cdot 2^{2k-1}}{2^{6k} + 2^{5k+1}} \in \mathbb{SQ}.$$

- $X = 2^k$, so that $f(k) = \tilde{f}(2^k)$, where

$$\tilde{f}(X) = \frac{X^3 + 5/2X^2}{X^6 + 2X^5} \in \mathbb{Z}(X).$$

$$\mathbf{p} = \mathbf{2k} \quad \text{and} \quad \mathbf{f}(\mathbf{k}) = \tilde{\mathbf{f}}(\mathbf{2}^{\mathbf{k}}).$$

$$f(k) = \frac{2^{3k} + 5 \cdot 2^{2k-1}}{2^{6k} + 2^{5k+1}}$$

Asymptotic behavior as $k \to \infty$:

$f(k) \sim 2^{-3k}$

$\hookrightarrow 2^{-3k-1} \leqslant f(k) < 2^{-3k+1}$,

$$\tilde{f}(X) = \frac{2X + 5}{2X^4 + 4X^3}$$

Asymptotic behavior as $X \to \infty$:

$\tilde{f}(X) \sim X^{-3}$

$\hookrightarrow \frac{1}{2}X^{-3} \leqslant \tilde{f}(X) < 2X^{-3}$,

$$\mathbf{p} = \mathbf{2k} \quad \text{and} \quad \mathbf{f}(\mathbf{k}) = \tilde{\mathbf{f}}(\mathbf{2^k}).$$

$$f(k) = \frac{2^{3k} + 5 \cdot 2^{2k-1}}{2^{6k} + 2^{5k+1}}$$

Asymptotic behavior as $k \to \infty$:

$$f(k) = 2^{-3k} + 2^{-2k-1} + \mathcal{O}(2^{-5k})$$

$$\hookrightarrow \; 2^{-3k} \leqslant f(k) < 2^{-3k+1},$$

$$\tilde{f}(X) = \frac{2X + 5}{2X^4 + 4X^3}$$

Asymptotic behavior as $X \to \infty$:

$$\tilde{f}(X) = X^{-3} + \tfrac{1}{2}X^{-4} + \mathcal{O}(X^{-5})$$

$$\hookrightarrow X^{-3} \leqslant \tilde{f}(X) < 2X^{-3},$$

$\text{exponent}(f) = -3k$ and $\text{ulp}_p(f) = 2^{-5k+1} \leftrightarrow 2X^{-5}$, for all $k \in \mathbb{N}^*$.

Rounding to the nearest integer: $\text{RN}_p(f) = \text{ulp}_p(f) \cdot \Big\lfloor \underbrace{\frac{f}{\text{ulp}_p(f)}}_{g} \Big\rceil$

$$\mathbf{RN_p(f) = 2^{-5k+1} \cdot \lfloor g \rceil \quad \text{and} \quad g(k) = \tilde{g}(2^k).}$$

First step toward $\lfloor g \rceil$: find a **symbolic integer** that **approximates** g.

$$g(k) = \frac{2^{3k} + 5 \cdot 2^{2k-1}}{2^{k+1} + 2^2}$$

Asymptotic behavior as $k \to \infty$:

$$g(k) = \underbrace{2^{2k-1} + 2^{k-2}}_{h(k)} + \mathcal{O}(1)$$

$$\tilde{g}(X) = \frac{2X^3 + 5X^2}{4X + 8}$$

Laurent expansion as $X \to \infty$:

$$\tilde{g}(X) = \underbrace{\frac{1}{2}X^2 + \frac{1}{4}X}_{\tilde{h}(k)} + \mathcal{O}(1)$$

- $h(k) \in \mathbb{Z}$, for all $k \geqslant 2$: h is a **symbolic integer**;
- $g(k) - h(k) = \mathcal{O}(1)$ as $k \to \infty$: h **approximates** g.

$\mathrm{RN}_{\mathbf{p}}(\mathbf{f}) = \mathbf{2^{-5k+1}} \cdot \lfloor \mathbf{g} \rceil$ and h **symbolic integer approximating** g.

Second step to $\lfloor g \rceil$: **correct** h if needed.

$$h(k) = 2^{2k-1} + 2^{k-2}$$

Distance to g, as $k \to \infty$:

$$|g(k) - h(k)| = \frac{2^k}{2^{k+1} + 2^2} < 1/2$$

$$\tilde{h}(X) = \frac{1}{2}X^2 + \frac{1}{4}X$$

Distance to $\tilde{g}$, as $X \to \infty$:

$$|\tilde{g}(X) - \tilde{h}(X)| = \frac{X}{2X+4} < 1/2$$

$\hookrightarrow \lfloor g(k) \rceil = h(k)$, for all $k \geqslant 2$.

$$\mathrm{RN}_p(f) = 2^{-3k} + 2^{-4k-1}, \text{ for } k \geqslant 2, \text{ with } p = 2k.$$

Comparing with the earlier numerical computations:

$\vdots$

$\mathrm{RN}_{24}(f(24)) = 2^{-36} + 2^{-49},$ [ok]

$\vdots$

$\mathrm{RN}_{6}(f(6)) = 2^{-9} + 2^{-13},$ [ok]

$\mathrm{RN}_{4}(f(4)) = 2^{-6} + 2^{-9},$ [ok]

$\mathrm{RN}_{2}(f(2)) = 2^{-3},$ [ko]

$\hookrightarrow$ Our computation matches the classical ones for all $k \geqslant 2$.

We define the following functions on $\mathbb{SQ}$:

- **sign**,
- $\hookrightarrow$ **comparisons** and **absolute value**,
- $\hookrightarrow$ **exponent**,
- $\hookrightarrow$ $\mathbf{ulp}_p$.

Asymptotic behaviors, with a domain of validity ($k \geqslant k_0$).

We define the following functions on $\mathbb{SQ}$:

- **sign**,

$\hookrightarrow$ **comparisons** and **absolute value**,

$\hookrightarrow$ **exponent**,

$\hookrightarrow$ **ulp**$_p$.

Asymptotic behaviors, with a domain of validity ($k \geqslant k_0$).

$$\mathsf{RN}_p : \mathbb{SQ} \to \mathbb{SF}_p \quad \leftrightarrow \quad \lfloor \cdot \rceil : \mathbb{SQ} \to \mathbb{SZ}$$

but

Some elements of $\mathbb{SQ}$ cannot be rounded.

Outline

In base 2 and precision $p = k$, consider [BPZ07]

$$f(k) = \frac{2}{3}(1 + 11 \cdot 2^{-k})$$

Does $f \in \mathbb{SF}_p$?

In base 2 and precision $p = k$, consider [BPZ07]

$$f(k) = \frac{2}{3}(1 + 11 \cdot 2^{-k})$$

Does $f \in \mathbb{SF}_p$?

$$\mathrm{ulp}_p(f) = 2^{-k} \quad \Rightarrow \quad g(k) = \frac{f}{\mathrm{ulp}_p(f)} = \frac{2}{3}(2^k + 11)$$

We have:

$$2^{k-1} \leqslant g(k) < 2^k \quad (\text{for } k \geqslant 5).$$

Does $g \in \mathbb{SZ}$?

In base 2 and precision $p = k$, consider [BPZ07]

$$f(k) = \frac{2}{3}(1 + 11 \cdot 2^{-k})$$

Does $f \in \mathbb{SF}_p$?

$$\mathrm{ulp}_p(f) = 2^{-k} \quad \Rightarrow \quad g(k) = \frac{f}{\mathrm{ulp}_p(f)} = \frac{2}{3}(2^k + 11)$$

We have:

$$2^{k-1} \leqslant g(k) < 2^k \quad (\text{for } k \geqslant 5).$$

Does $g \in \mathbb{SZ}$?

For $k \in \mathbb{N}$, we have $2^k + 11 \equiv (-1)^k - 1 \pmod 3$ so that:

- if k is even, then $g(k) \in \mathbb{Z}$;
- if k is odd, then $g(k) \notin \mathbb{Z}$.

$$g \notin \mathbb{SZ} \Rightarrow f \notin \mathbb{SF}_p; \quad \mathrm{RN}_p(f) = 2^{-k} \cdot \lfloor g \rceil; \quad \lfloor g \rceil = ?$$

$$g(k) = \frac{2}{3}(2^k + 11) \notin \mathbb{SZ} \quad \text{and} \quad g(k) \in \mathbb{Z} \text{ iff } k \text{ even.}$$

There is no **symbolic integer** that is the **nearest** to g.

Sketch of the proof (by **contradiction**): suppose $h \in \mathbb{SZ}$ **approximates** to g,

- $h(2k)$ is also a **symbolic integer** that **approximates** $g(2k)$;
- we saw that $g(2k) \in \mathbb{SZ}$;

$\hookrightarrow$ $h(2k) = g(2k) + n$, with $n \in \mathbb{Z}$;

$\hookrightarrow$ $h = g + n$;

$\hookrightarrow$ $g = h - n \in \mathbb{SZ}$.

$$g(k) = \frac{2}{3}(2^k + 11) \notin \mathbb{SZ} \quad \text{and} \quad g(k) \in \mathbb{Z} \text{ iff } k \text{ even.}$$

There is no **symbolic integer** that is the **nearest** to g.

Sketch of the proof (by **contradiction**): suppose $h \in \mathbb{SZ}$ **approximates** to g,

- $h(2k)$ is also a **symbolic integer** that **approximates** $g(2k)$;
- we saw that $g(2k) \in \mathbb{SZ}$;

$\hookrightarrow$ $h(2k) = g(2k) + n$, with $n \in \mathbb{Z}$;

$\hookrightarrow$ $h = g + n$;

$\hookrightarrow$ $g = h - n \in \mathbb{SZ}$.

We cannot round f to a nearest symbolic floating-point number but:

- $f(2k) \in \mathbb{SF}_p$ for $p = 2k$;
- $\mathrm{RN}_p(f(2k+1)) = (2^{2k+2} + 23)/3$ for $p = 2k+1$.

Outline

```
> Kahan := proc (a, b, c, d, p)
  local wh, e, fh;
   wh := rn(b·c, p);
   e := wh - b·c;
   fh := rn(a·d - wh, p);
   rn(fh + e, p)
  end proc:
> p := 2·k:
  a := SQ(2^(p-1) + 1); b := SQ(2^(p-1) + 1) : c := SQ(2^(p-1) + 2^(p-2)) : d := SQ(2^p + 2^(p-2)) :
```

$$a := \frac{1}{2}\,4^k + 1$$

```
> x := a·d − b·c;
  xh := Kahan(a, b, c, d, p);
  SQ:-getW(xh);
```

$$x := \frac{1}{4}\,16^k + \frac{1}{2}\,4^k$$

$$xh := \frac{1}{4}\,16^k$$

$$1$$

```
> err := abs((x − xh)/x);
  simplify(toU(err, p, u));
```

$$err := \frac{2}{4^k + 2}$$

$$\frac{2\,u}{1 + 2\,u}$$

$> p := 2\,k;$

$$f := SQ\left(\frac{2^{\frac{3\cdot p}{2}} + 5\cdot 2^{p-1}}{2^{3\cdot p} + 2^{\frac{5\cdot p}{2}+1}}\right);$$

$fh := rn(f, p);$

$SQ\text{:-}getk0(fh);$

$>$

$$p := 2\,k$$

$$f := \frac{1}{2}\,\frac{2\,8^k + 5\,4^k}{64^k + 2\,32^k}$$

$$fh := 8^{-k} + \frac{1}{2}\,16^{-k}$$

$$3$$

```
> p := k;
  f := SQ(2/3·(1 + 11·2^(-p)));
  fh := rn(f, p);
  SQ:-getk0(fh);
  SQ:-getW(fh);
```

$$p := k$$

$$f := \frac{2}{3} + \frac{22}{3}\, 2^{-k}$$

$$fh := \frac{2}{3} + \frac{22}{3}\, 2^{-k}$$

$$6$$

$$2$$

```
> p := 2·k + 1;
  f := SQ(2/3·(1 + 11·2^(-p)));
  fh := rn(f, p);
  SQ:-getk0(fh);
  SQ:-getW(fh);
```

$$p := 2\,k + 1$$

$$f := \frac{2}{3} + \frac{11}{3}\, 4^{-k}$$

$$fh := \frac{2}{3} + \frac{23}{6}\, 4^{-k}$$

$$3$$

$$1$$

Conclusion and perspectives

The current library:

- rigorous formalism for symbolic floating-point arithmetic;
- effective implementation in Maple:
  27 examples [BPZ07, JLM13a, JLM13b, Mul15] within 1.5s on this laptop;
- other rounding modes are implemented.

Preprint available at `https://hal.inria.fr/hal-01232159`

Perspectives

- extend the model to handle more operations;
- automatic search for examples for which the final error is close to the bound;
- transfer to a formal proof system to increase the confidence.

References

A. Avizienis.
Signed-digit number representations for fast parallel arithmetic.
IRE Transactions on Electronic Computers, 10:389–400, 1961.

N. Brisebarre and J.-M. Muller.
Correct rounding of algebraic functions.
Theoretical Informatics and Applications, 41:71–83, 2007.

Richard Brent, Colin Percival, and Paul Zimmermann.
Error bounds on complex floating-point multiplication.
Mathematics of Computation, 76:1469–1481, 2007.

IEEE Computer Society.
IEEE Standard for Floating-Point Arithmetic.
IEEE Standard 754-2008, August 2008.
available at http://ieeexplore.ieee.org/servlet/opac?punumber=4610933.

References

C. Iordache and D. W. Matula.
On infinitely precise rounding for division, square root, reciprocal and square root reciprocal.
In *14th IEEE Symposium on Computer Arithmetic, Adelaide, Australia*, pages 233–240, 1999.

Claude-Pierre Jeannerod, Nicolas Louvet, and Jean-Michel Muller.
Further analysis of Kahan's algorithm for the accurate computation of 2×2 determinants.
Mathematics of Computation, 82:2245–2264, 2013.

Claude-Pierre Jeannerod, Nicolas Louvet, and Jean-Michel Muller.
On the componentwise accuracy of complex floating-point division with an FMA.
In *21st IEEE Symposium on Computer Arithmetic, Austin, TX, USA*, pages 83–90, 2013.

Jean-Michel Muller.
On the error of computing $ab + cd$ using Cornea, Harrison and Tang's method.
ACM Transactions on Mathematical Software, 41(2):7:1–7:8, February 2015.