Formal proofs and computations in finite precision arithmetic

Sylvie Boldo* & Marc Daumas* & Laurent Théry**

École normale de Lyon, France*

Università di L'Aquila, Italy**

Cycle

[1996] First meeting at Rome

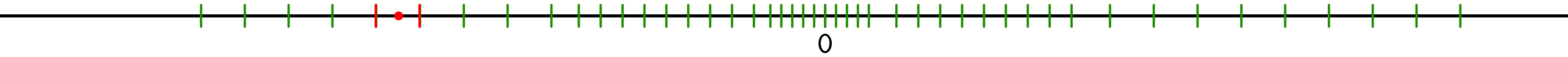
[1997] Buchberger's Algorithm

...

[1999] AOC initiative

[2003] Back to Rome

Floating-point numbers



Real numbers $\mathbb{R}$

Floating point numbers $\mathbb{F}$: $m \times \beta^e$

Rounding $\circ$

IEEE-754 standard

$$a \oplus b = \circ(a + b)$$

Proving Basic Properties

Sterbenz Theorem

$\forall p, q : \mathbb{F}.$

$$q/2 \leq p \leq 2q \Rightarrow p \ominus q = p - q$$

Proving Basic Properties

Malcolm Program

```
x := 1.0;
y := 1.0;
while ((x + 1.0) - x) = 1.0 do
   x := 2.0 * x;
while ((x + y) - x) <> y do
   y := y + 1.0;
```

This program always ends and y contains β

Proving New Properties

Axpy_opt Theorem

Given real numbers a_0, x_0 and y_0, and floating point numbers a, x and y, if no overflow occurs and if

$$5\,\frac{2+\mathrm{ulp}}{2-\mathrm{ulp}} \times \left(|a \times x| + \frac{\lambda \times \mathrm{ulp}}{2}\right) \leq |y| \quad \text{and}$$

$$|y_0 - y| + |a_0 \times x_0 - a \times x| \leq \frac{\mathrm{ulp}}{8}\left((1-\mathrm{ulp}) \times |y| - |a \times x| - 2 \times \lambda\right),$$

then

$\circ(\circ(a \times x) + y)$ is a faithful approximation of $a_0 \times x_0 + y_0$.

Massaging inequalities

$$|p| \leq \frac{|z|}{1-\beta^{-\mathsf{prec}}} + \frac{\beta^{-e_{min}-1}}{1-\beta^{-\mathsf{prec}}}$$

$$\left(1-\beta^{-\mathsf{prec}}\right) \times |p| \leq |z| + \beta^{-e_{min}-1}$$

$$|p| - \left(|p| \times \beta^{-\mathsf{prec}} + \beta^{-e_{min}-1}\right) \leq |z|$$

$$|p| - \left(|p| \times \beta^{-\mathsf{prec}} + \beta^{-e_{min}-1}\right) \leq |p| - \frac{\mathrm{ulp}(p)}{\beta} \leq |z|$$

Simplifications

$$x(z+y+2) < y(3x+1)$$

$$\alpha x(\beta z+\delta y+\gamma) < \varepsilon y(\eta x+\iota)$$

$$\left\{\begin{array}{rcr} \alpha\beta & = & 1 \\ \alpha\delta-\varepsilon\eta & = & -2 \\ \alpha\gamma & = & 2 \\ \varepsilon\iota & = & 1 \end{array}\right.$$

Further applications

Deriving inequalities from function analysis

Numerical Programs

Filters

Numerical Linear Algebra

Function analysis

Elementary functions for $\mathbb{F}$

$f(x)$ approximated by a $P(X)$ on I

$$||f(x) - P(x)||_\infty \leq \varepsilon$$

First Technique

Taylor Expansion

$$||f(x) - P(x)||_\infty \leq ||f(x) - T(x)||_\infty + ||T(x) - P(x)||_\infty$$

$$||f(x) - P(x)||_\infty \leq \varepsilon_1 + ||T(x) - P(x)||_\infty$$

Sturm's series

$$||f(x) - P(x)||_\infty \leq \varepsilon_1 + \varepsilon_2$$

$\Longrightarrow$ exp function (J. Harrison)

Second Technique

Interval arithmetic

$$x \in I \Rightarrow g(x) \in g(I)$$

Example:

$$x \in [a_1, b_1] \wedge y \in [a_2, b_2] \Rightarrow x + y \in [a_1 \oplus a_2, b_1 \oplus b_2]$$

Pessimistic bound

Second Technique

Interval arithmetic + Bisection

Init: $\Phi \leftarrow \{I\}$, $f_{max} \leftarrow -\infty$

Iteration:

Pick $J \in \Phi$, $\Phi = \Phi - \{J\}$ and $f(J) = [min, max]$

If $f_{max} < max$ then

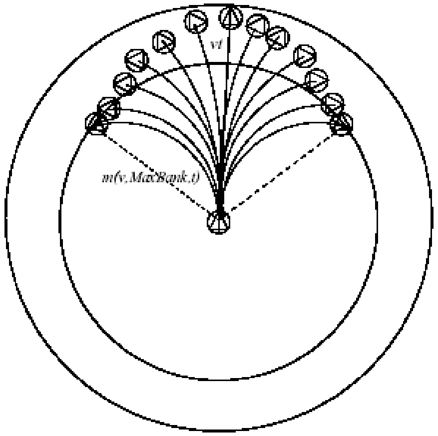
$$\Phi = \Phi + \{J^{1/2-}, J^{1/2+}\}$$
$$f_{max} = \mathtt{max}(f_{max}, min)$$

If J too small, $f_{max} = bound$

Numerical Programs

Complete verification of numerical programs

Example: Conflict Detection



Formal Verification for $\mathbb{R}$ (Muñoz et al)
Proof for $\mathbb{F}$ (Melquiond)

Filters

Function with result in {True,False,Don't Know}

Example: Computational Geometry

Certification of the result

Adaptative algorithms

Numerical Linear Algebra

Unfortunately too ambitious for the moment

Conclusions

Few achieved, Lots to be done

Need for both proofs and computations

How to?

- External CAS: Reliability?
- Internal Computation: 1 second
- Oracles: Generality?
- Certified Computation