Floating Point Arithmetic and Rounding Error Analysis

Claude-Pierre Jeannerod – Nathalie Revol

Inria – LIP, ENS de Lyon

Context

Starting point:

How do numerical algorithms behave in finite precision arithmetic?

Typically,

- basic matrix computations: $Ax = b$, ...
- floating-point arithmetic as specified by IEEE 754.

Context

Starting point:

How do numerical algorithms behave in finite precision arithmetic?

Typically,

- basic matrix computations: $Ax = b$, ...
- floating-point arithmetic as specified by IEEE 754.

Finite precision $\quad\Rightarrow\quad$ rounding errors:

$$2.3456 \times 5.4321 = 12.74153376$$

Context

Starting point:

How do numerical algorithms behave in finite precision arithmetic?

Typically,

- basic matrix computations: $Ax = b$, ...
- floating-point arithmetic as specified by IEEE 754.

Finite precision $\Rightarrow$ rounding errors:

$$2.3456 \times 5.4321 = 12.74153376$$
$$2.3456 \;/\; 5.4321 = 0.431803538226468584\ldots$$

Context

Starting point:

How do numerical algorithms behave in finite precision arithmetic?

Typically,

- basic matrix computations: $Ax = b$, ...
- floating-point arithmetic as specified by IEEE 754.

Finite precision $\Rightarrow$ rounding errors:

$$2.3456 \times 5.4321 = 12.74153376$$
$$2.3456 \; / \; 5.4321 = 0.43180353822646858489...$$
$$2.3456 + 5.4321 = 7.7777$$

Context

Starting point:

How do numerical algorithms behave in finite precision arithmetic?

Typically,

- basic matrix computations: $Ax = b$, ...
- floating-point arithmetic as specified by IEEE 754.

Finite precision $\Rightarrow$ rounding errors:

$$2.3456 \times 5.4321 = 12.7415\color{red}{3376}$$
$$2.3456 \;/\; 5.4321 = 0.43180\color{red}{35382264685848 9\ldots}$$
$$2.3456 + 5.4321 = 7.7777$$

What is the effect of all such errors on the computed solution $\widehat{x}$?

Rounding error analysis

Old and nontrivial question [von Neumann, Turing, Wilkinson, ...]

Rounding error analysis

Old and nontrivial question [von Neumann, Turing, Wilkinson, ...]

In this lecture, two approaches:

A priori analysis:

- Goal: bound on $\|\widehat{x} - x\| / \|x\|$ for any input and format
- Tool: the many nice properties of floating-point
- Ideal: readable, provably tight bound + short proof

Rounding error analysis

Old and nontrivial question [von Neumann, Turing, Wilkinson, ...]

In this lecture, two approaches:

A priori analysis:

- Goal: bound on $\|\widehat{x} - x\|/\|x\|$ for any input and format
- Tool: the many nice properties of floating-point
- Ideal: readable, provably tight bound + short proof

A posteriori, automatic analysis:

- Goal: $\widehat{x}$ and enclosure of $\widehat{x} - x$ for given input and format
- Tool: interval arithmetic based on floating-point
- Ideal: a narrow interval computed fast

Rounding error analysis

Old and nontrivial question [von Neumann, Turing, Wilkinson, ...]

In this lecture, two approaches:

A priori analysis: $\rightarrow$ this lecture

- Goal: bound on $\|\widehat{x} - x\| / \|x\|$ for any input and format
- Tool: the many nice properties of floating-point
- Ideal: readable, provably tight bound + short proof

A posteriori, automatic analysis: $\rightarrow$ Nathalie's lecture

- Goal: $\widehat{x}$ and enclosure of $\widehat{x} - x$ for given input and format
- Tool: interval arithmetic based on floating-point
- Ideal: a narrow interval computed fast

Context

Floating-point arithmetic

A priori analysis

Conclusion

Floating-point arithmetic

- An approximation of arithmetic over $\mathbb{R}$.

- 1940's: first implementations [Zuse's computers].
- 1985-2008: full specification [IEEE 754 standard].
- Today: IEEE arithmetic everywhere!

Floating-point arithmetic

- An approximation of arithmetic over $\mathbb{R}$.

- 1940's: first implementations [Zuse's computers].
- 1985-2008: full specification [IEEE 754 standard].
- Today: IEEE arithmetic everywhere!

- Although often considered as fuzzy, it is highly structured and has many nice mathematical properties.

Floating-point arithmetic

- An approximation of arithmetic over $\mathbb{R}$.

- 1940's: first implementations [Zuse's computers].
- 1985-2008: full specification [IEEE 754 standard].
- Today: IEEE arithmetic everywhere!

- Although often considered as fuzzy, it is highly structured and has many nice mathematical properties.

$\hookrightarrow$ **How to exploit these properties for rigorous analyses?**

Floating-point numbers

Rational numbers of the form $M \cdot \beta^E$, where

$$M, E \in \mathbb{Z}, \qquad |M| < \beta^p, \qquad E + p - 1 \in [e_{\min}, e_{\max}].$$

- base β,
- precision p,
- exponent range defined by $e_{\min}$ and $e_{\max}$.

Floating-point numbers

Rational numbers of the form $M \cdot \beta^E$, where

$$M, E \in \mathbb{Z}, \qquad |M| < \beta^p, \qquad E + p - 1 \in [e_{\min}, e_{\max}].$$

- base β,
- precision p,
- exponent range defined by $e_{\min}$ and $e_{\max}$.

Floats in C have $\beta = 2$, $p = 24$, and $[e_{\min}, e_{\max}] = [-126, 127]$.

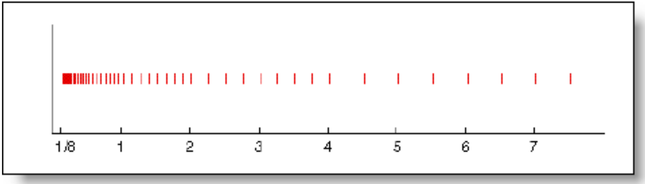
Floating-point numbers

Rational numbers of the form $M \cdot \beta^E$, where

$$M, E \in \mathbb{Z}, \qquad |M| < \beta^p, \qquad E + p - 1 \in [e_{\min}, e_{\max}].$$

- base β,
- precision p,
- exponent range defined by $e_{\min}$ and $e_{\max}$.

Floats in C have $\beta = 2$, $p = 24$, and $[e_{\min}, e_{\max}] = [-126, 127]$.



Floating-point numbers

We assume

- $e_{\min} = -\infty$ and $e_{\max} = +\infty$: unbounded exponent range,
- β is even and $p \geqslant 2$.

Floating-point numbers

We assume

- $e_{\min} = -\infty$ and $e_{\max} = +\infty$: unbounded exponent range,
- β is even and $p \geqslant 2$.

Definition: The set $\mathbb{F}$ of floating-point numbers in base β and precision p is

$$\mathbb{F} := \{0\} \cup \left\{ M \cdot \beta^E \ : \ M, E \in \mathbb{Z}, \quad \beta^{p-1} \leqslant |M| < \beta^p \right\}.$$

Floating-point numbers: other representations

- $x \in \mathbb{F} \backslash \{0\} \quad \Rightarrow \quad |x| = m \cdot \beta^e, \quad m = (* \bullet \underbrace{* \cdots *}_{p-1})_\beta \in [1, \beta).$

- Three useful "units":
  - Unit in the first place: $\mathrm{ufp}(x) = \beta^e$,
  - Unit in the last place: $\mathrm{ulp}(x) = \beta^{e-p+1}$,
  - Unit roundoff: $u = \frac{1}{2}\beta^{1-p}$.

Floating-point numbers: other representations

- $x \in \mathbb{F}\backslash\{0\} \quad \Rightarrow \quad |x| = m \cdot \beta^e, \quad m = (*\bullet\underbrace{*\cdots*}_{p-1})_\beta \in [1, \beta).$

- Three useful "units":
  - Unit in the first place: $\mathrm{ufp}(x) = \beta^e$,
  - Unit in the last place: $\mathrm{ulp}(x) = \beta^{e-p+1}$,
  - Unit roundoff: $u = \frac{1}{2}\beta^{1-p}$.

- Alternative views, which display the structure of $\mathbb{F}$ very well:
  - $x \in \mathrm{ulp}(x)\mathbb{Z}$,
  - $|x| = (1 + 2ku)\,\mathrm{ufp}(x), \quad k \in \mathbb{N}.$

$$\Rightarrow \quad \mathbb{F} \cap [1, \beta) = \Big\{1, 1 + 2u, 1 + 4u, \ldots\Big\}.$$

Floating-point numbers: some properties

$\mathbb{F}$ can be seen as a structured grid with many nice properties:

- Symmetry: $f \in \mathbb{F} \quad \Rightarrow \quad -f \in \mathbb{F}$;
- Auto-similarity: $f \in \mathbb{F},\ e \in \mathbb{Z} \quad \Rightarrow \quad f \cdot \beta^e \in \mathbb{F}$;

Floating-point numbers: some properties

$\mathbb{F}$ can be seen as a structured grid with many nice properties:

- Symmetry: $f \in \mathbb{F} \quad \Rightarrow \quad -f \in \mathbb{F}$;
- Auto-similarity: $f \in \mathbb{F},\ e \in \mathbb{Z} \quad \Rightarrow \quad f \cdot \beta^e \in \mathbb{F}$;
- $\mathbb{F} \cap [1, \beta) = \left\{1, 1 + 2u, 1 + 4u, \ldots\right\}$;

Floating-point numbers: some properties

$\mathbb{F}$ can be seen as a structured grid with many nice properties:

- Symmetry: $f \in \mathbb{F} \quad \Rightarrow \quad -f \in \mathbb{F}$;
- Auto-similarity: $f \in \mathbb{F},\ e \in \mathbb{Z} \quad \Rightarrow \quad f \cdot \beta^e \in \mathbb{F}$;
- $\mathbb{F} \cap [1, \beta) = \left\{ 1, 1 + 2u, 1 + 4u, \dots \right\}$;
- $\mathbb{F} \cap [\beta^e, \beta^{e+1}]$ has $(\beta - 1)\beta^{p-1} + 1$ equally spaced elements, with spacing equal to
$$2u\beta^e\,;$$

Floating-point numbers: some properties

$\mathbb{F}$ can be seen as a structured grid with many nice properties:

- Symmetry: $f \in \mathbb{F} \quad \Rightarrow \quad -f \in \mathbb{F}$;
- Auto-similarity: $f \in \mathbb{F},\ e \in \mathbb{Z} \quad \Rightarrow \quad f \cdot \beta^e \in \mathbb{F}$;
- $\mathbb{F} \cap [1, \beta) = \Big\{1, 1 + 2u, 1 + 4u, \dots\Big\}$;
- $\mathbb{F} \cap [\beta^e, \beta^{e+1}]$ has $(\beta - 1)\beta^{p-1} + 1$ equally spaced elements, with spacing equal to
$$2u\beta^e\,;$$
- Neighborhood of $1 \in \mathbb{F}$:
$$\dots, 1 - \tfrac{4u}{\beta},\ 1 - \tfrac{2u}{\beta},\ 1,\ 1 + 2u,\ 1 + 4u, \dots.$$

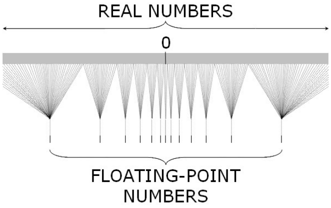
Hence 1 is β x closer to its predecessor than to its successor.

Rounding functions

Round-to-nearest function $\mathrm{RN} : \mathbb{R} \to \mathbb{F}$ such that

$$\forall t \in \mathbb{R}, \qquad |\mathrm{RN}(t) - t| = \min_{f \in \mathbb{F}} |f - t|,$$

with given tie-breaking rule.

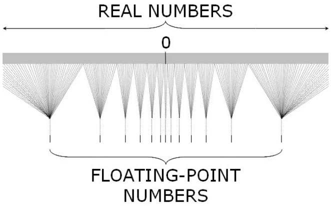


Rounding functions

Round-to-nearest function RN : $\mathbb{R} \to \mathbb{F}$ such that

$$\forall t \in \mathbb{R}, \qquad |\mathrm{RN}(t) - t| = \min_{f \in \mathbb{F}} |f - t|,$$

with given tie-breaking rule.



- $t \in \mathbb{F} \;\Rightarrow\; \mathrm{RN}(t) = t$
- RN nondecreasing
- reasonable tie-breaking rule:
  - $\mathrm{RN}(-t) = -\mathrm{RN}(t)$
  - $\mathrm{RN}(t\beta^e) = \mathrm{RN}(t)\beta^e, \;\; e \in \mathbb{Z}$

Rounding functions

More generally, a rounding function $\circ$ is any map from $\mathbb{R}$ to $\mathbb{F}$ such that

$$t \in \mathbb{F} \quad \Rightarrow \quad \circ(t) = t; \qquad\qquad t \leqslant t' \quad \Rightarrow \quad \circ(t) \leqslant \circ(t').$$

Rounding functions

More generally, a rounding function $\circ$ is any map from $\mathbb{R}$ to $\mathbb{F}$ such that

$$t \in \mathbb{F} \quad \Rightarrow \quad \circ(t) = t; \qquad\qquad t \leqslant t' \quad \Rightarrow \quad \circ(t) \leqslant \circ(t').$$

Adding just one extra constraint gives the usual directed roundings:

- Rounding down: $\mathrm{RD}(t) \leqslant t$.
- Rounding up: $t \leqslant \mathrm{RU}(t)$.
- Rounding to zero: $|\mathrm{RZ}(t)| \leqslant |t|$.

Rounding functions

More generally, a rounding function $\circ$ is any map from $\mathbb{R}$ to $\mathbb{F}$ such that

$$t \in \mathbb{F} \quad \Rightarrow \quad \circ(t) = t; \qquad\qquad t \leqslant t' \quad \Rightarrow \quad \circ(t) \leqslant \circ(t').$$

Adding just one extra constraint gives the usual directed roundings:

- Rounding down: $\mathrm{RD}(t) \leqslant t$.
- Rounding up: $t \leqslant \mathrm{RU}(t)$.
- Rounding to zero: $|\mathrm{RZ}(t)| \leqslant |t|$.

Key property for interval arithmetic:

$$t \notin \mathbb{F} \qquad \Rightarrow \qquad t \in \Big[\mathrm{RD}(t), \mathrm{RU}(t)\Big].$$

Error bounds for real numbers

$$E_1(t) := \frac{|\mathrm{RN}(t) - t|}{|t|} \leqslant \frac{u}{1+u}, \qquad E_2(t) := \frac{|\mathrm{RN}(t) - t|}{|\mathrm{RN}(t)|} \leqslant u.$$

Error bounds for real numbers

$$E_1(t) := \frac{|\mathrm{RN}(t) - t|}{|t|} \leqslant \frac{u}{1+u}, \qquad E_2(t) := \frac{|\mathrm{RN}(t) - t|}{|\mathrm{RN}(t)|} \leqslant u.$$

Proof:

- Assume $1 \leqslant t < \beta$, so that

$$\mathrm{RN}(t) \in \{1, 1+2u, 1+4u, \ldots, \beta\}.$$

Error bounds for real numbers

$$E_1(t) := \frac{|\mathrm{RN}(t) - t|}{|t|} \leqslant \frac{u}{1+u}, \qquad E_2(t) := \frac{|\mathrm{RN}(t) - t|}{|\mathrm{RN}(t)|} \leqslant u.$$

Proof:

- Assume $1 \leqslant t < \beta$, so that

$$\mathrm{RN}(t) \in \{1, 1+2u, 1+4u, \ldots, \beta\}.$$

- Then $|\mathrm{RN}(t) - t| \leqslant \frac{1}{2} \times 2u = u$.

Error bounds for real numbers

$$E_1(t) := \frac{|\mathrm{RN}(t) - t|}{|t|} \leqslant \frac{u}{1+u}, \qquad E_2(t) := \frac{|\mathrm{RN}(t) - t|}{|\mathrm{RN}(t)|} \leqslant u.$$

Proof:

- Assume $1 \leqslant t < \beta$, so that

$$\mathrm{RN}(t) \in \{1, 1+2u, 1+4u, \ldots, \beta\}.$$

- Then $|\mathrm{RN}(t) - t| \leqslant \frac{1}{2} \times 2u = u$.
- Dividing by $\mathrm{RN}(t) \geqslant 1$ gives directly the bound on E_2.

Error bounds for real numbers

$$E_1(t) := \frac{|\mathrm{RN}(t) - t|}{|t|} \leqslant \frac{u}{1+u}, \qquad E_2(t) := \frac{|\mathrm{RN}(t) - t|}{|\mathrm{RN}(t)|} \leqslant u.$$

Proof:

- Assume $1 \leqslant t < \beta$, so that

$$\mathrm{RN}(t) \in \{1, 1+2u, 1+4u, \dots, \beta\}.$$

- Then $|\mathrm{RN}(t) - t| \leqslant \frac{1}{2} \times 2u = u$.
- Dividing by $\mathrm{RN}(t) \geqslant 1$ gives directly the bound on E_2.
- If $t \geqslant 1 + u$ then the bound $E_1(t) \leqslant \frac{u}{1+u}$ follows.

Error bounds for real numbers

$$E_1(t) := \frac{|\mathrm{RN}(t) - t|}{|t|} \leqslant \frac{u}{1+u}, \qquad E_2(t) := \frac{|\mathrm{RN}(t) - t|}{|\mathrm{RN}(t)|} \leqslant u.$$

Proof:

- Assume $1 \leqslant t < \beta$, so that

$$\mathrm{RN}(t) \in \{1, 1+2u, 1+4u, \ldots, \beta\}.$$

- Then $|\mathrm{RN}(t) - t| \leqslant \frac{1}{2} \times 2u = u$.
- Dividing by $\mathrm{RN}(t) \geqslant 1$ gives directly the bound on E_2.
- If $t \geqslant 1+u$ then the bound $E_1(t) \leqslant \frac{u}{1+u}$ follows.
- Else $1 \leqslant t < 1+u \;\Rightarrow\; \mathrm{RN}(t) = 1 \;\Rightarrow\; E_1(t) = \frac{t-1}{t} < \frac{u}{1+u}$. $\square$

Error bounds for real numbers

$$E_1(t) := \frac{|\mathrm{RN}(t) - t|}{|t|} \leqslant \frac{u}{1+u}, \qquad E_2(t) := \frac{|\mathrm{RN}(t) - t|}{|\mathrm{RN}(t)|} \leqslant u.$$

Proof:

- Assume $1 \leqslant t < \beta$, so that

$$\mathrm{RN}(t) \in \{1, 1+2u, 1+4u, \ldots, \beta\}.$$

- Then $|\mathrm{RN}(t) - t| \leqslant \frac{1}{2} \times 2u = u$.
- Dividing by $\mathrm{RN}(t) \geqslant 1$ gives directly the bound on E_2.
- If $t \geqslant 1 + u$ then the bound $E_1(t) \leqslant \frac{u}{1+u}$ follows.
- Else $1 \leqslant t < 1 + u \;\Rightarrow\; \mathrm{RN}(t) = 1 \;\Rightarrow\; E_1(t) = \frac{t-1}{t} < \frac{u}{1+u}$. $\square$

Bound $\frac{u}{1+u}$: sharp and well known [Dekker'71, Holm'80, Knuth'81-98], but simpler bound u almost always used in practice.

Error bounds for real numbers

- Example for the usual binary formats:

$$u \approx \frac{u}{1+u} \approx \begin{cases} 4.9 \times 10^{-4} & \text{if } p = 11 \quad \text{(half)}, \\ 5.9 \times 10^{-8} & \text{if } p = 24 \quad \text{(float)}, \\ 1.1 \times 10^{-16} & \text{if } p = 53 \quad \text{(double)}, \\ 9.6 \times 10^{-35} & \text{if } p = 113 \quad \text{(quad)}. \end{cases}$$

- For directed roundings, replace these bounds by $2u$.

Conclusion: in all cases, the relative errors due to rounding can be bounded by a tiny quantity which depends only on the format.

Correct rounding

This is the result of the composition of two functions: basic operations performed exactly, and exact result then rounded:

$$x, y \in \mathbb{F}, \quad \mathrm{op} = \pm, \times, \div \qquad \Rightarrow \qquad \text{return } \hat{r} := \mathrm{RN}(x \,\mathrm{op}\, y).$$

op extends to square root and FMA (fused multiply add: $xy + z$).

Correct rounding

This is the result of the composition of two functions: basic operations performed exactly, and exact result then rounded:

$$x, y \in \mathbb{F}, \quad \mathrm{op} = \pm, \times, \div \qquad \Rightarrow \qquad \text{return } \widehat{r} := \mathrm{RN}(x \,\mathrm{op}\, y).$$

op extends to square root and FMA (fused multiply add: $xy + z$).

- The error bounds on E_1 and E_2 yield two standard models:

$$\begin{aligned} \widehat{r} &= (x \,\mathrm{op}\, y) \times (1 + \delta_1), \qquad & |\delta_1| \leqslant \tfrac{u}{1+u} =: u_1, \\ &= (x \,\mathrm{op}\, y) \times \frac{1}{1 + \delta_2}, \qquad & |\delta_2| \leqslant u. \end{aligned}$$

Example

Let $r = \frac{x+y}{2}$ be evaluated naively as $\widehat{r} = \mathrm{RN}\left(\frac{\mathrm{RN}(x+y)}{2}\right)$.

Example

Let $r = \frac{x+y}{2}$ be evaluated naively as $\widehat{r} = \mathrm{RN}\left(\frac{\mathrm{RN}(x+y)}{2}\right)$.

- High relative accuracy is ensured:

$$\begin{aligned}
\widehat{r} &= \frac{\mathrm{RN}(x+y)}{2}(1+\delta_1), \qquad |\delta_1| \leqslant u_1, \\
&= \frac{x+y}{2}(1+\delta_1)(1+\delta_1'), \qquad |\delta_1'| \leqslant u_1, \\
&=: r\,(1+\epsilon), \qquad |\epsilon| \leqslant 2u.
\end{aligned}$$

Example

Let $r = \frac{x+y}{2}$ be evaluated naively as $\widehat{r} = \mathrm{RN}\left(\frac{\mathrm{RN}(x+y)}{2}\right)$.

- High relative accuracy is ensured:

$$\begin{aligned}
\widehat{r} &= \frac{\mathrm{RN}(x+y)}{2}(1+\delta_1), \qquad |\delta_1| \leqslant u_1, \\
&= \frac{x+y}{2}(1+\delta_1)(1+\delta_1'), \qquad |\delta_1'| \leqslant u_1, \\
&=: r\,(1+\epsilon), \qquad |\epsilon| \leqslant 2u.
\end{aligned}$$

- We'd also like to have $\min(x,y) \leqslant \widehat{r} \leqslant \max(x,y)$ ...

Example

✗ Not always true:

$$\beta = 10,\ \ p = 3 \quad \Rightarrow \quad \mathrm{RN}\left(\frac{\mathrm{RN}\big(5.01 + 5.03\big)}{2}\right) = \mathrm{RN}\left(\frac{10}{2}\right) = 5.$$

Example

✗ Not always true:

$$\beta = 10,\ \ p = 3 \quad \Rightarrow \quad \mathrm{RN}\left(\frac{\mathrm{RN}\big(5.01 + 5.03\big)}{2}\right) = \mathrm{RN}\left(\frac{10}{2}\right) = 5.$$

✓ True if $\beta = 2$ or $\operatorname{sign}(x) \neq \operatorname{sign}(y)$.

Example

✗ Not always true:

$$\beta = 10,\ p = 3 \quad \Rightarrow \quad \mathrm{RN}\left(\frac{\mathrm{RN}(5.01 + 5.03)}{2}\right) = \mathrm{RN}\left(\frac{10}{2}\right) = 5.$$

✓ True if $\beta = 2$ or $\mathrm{sign}(x) \neq \mathrm{sign}(y)$.

Proof for base two:

- $\widehat{r} := \mathrm{RN}\left(\frac{\mathrm{RN}(x+y)}{2}\right) = \mathrm{RN}\left(\frac{x+y}{2}\right)$.
- $x \leqslant \frac{x+y}{2} \leqslant y \quad \Rightarrow \quad \mathrm{RN}(x) \leqslant \mathrm{RN}\left(\frac{x+y}{2}\right) \leqslant \mathrm{RN}(y)$

$$\Rightarrow \quad x \leqslant \widehat{r} \leqslant y. \qquad \square$$

Example

✗ Not always true:

$$\beta = 10,\ p = 3 \quad \Rightarrow \quad \mathrm{RN}\left(\frac{\mathrm{RN}(5.01 + 5.03)}{2}\right) = \mathrm{RN}\left(\frac{10}{2}\right) = 5.$$

✓ True if $\beta = 2$ or $\mathrm{sign}(x) \neq \mathrm{sign}(y)$.

Proof for base two:

- $\widehat{r} := \mathrm{RN}\left(\frac{\mathrm{RN}(x+y)}{2}\right) = \mathrm{RN}\left(\frac{x+y}{2}\right)$.
- $x \leqslant \frac{x+y}{2} \leqslant y \quad \Rightarrow \quad \mathrm{RN}(x) \leqslant \mathrm{RN}\left(\frac{x+y}{2}\right) \leqslant \mathrm{RN}(y)$

$$\Rightarrow \quad x \leqslant \widehat{r} \leqslant y. \qquad \square$$

↪ Repair other cases using $r = x + \frac{y-x}{2}$. [Sterbenz'74, Boldo'15]

Other typical floating-point surprises

- $\beta = 2$, any precision $p \quad \Rightarrow \quad \frac{1}{10} \notin \mathbb{F}$.

Other typical floating-point surprises

- $\beta = 2$, any precision p $\Rightarrow$ $\frac{1}{10} \notin \mathbb{F}$.

- Loss of algebraic properties: commutativity of $+$, $-$, $\times$ is preserved by correct rounding, but associativity and distributivity are lost:

$$\text{in general,} \quad \circ\left(\circ(x+y)+z\right) \neq \circ(x+\circ(y+z)).$$

Other typical floating-point surprises

- $\beta = 2$, any precision $p \quad \Rightarrow \quad \frac{1}{10} \notin \mathbb{F}$.
- Loss of algebraic properties: commutativity of $+$, $-$, $\times$ is preserved by correct rounding, but associativity and distributivity are lost:

$$\text{in general,} \quad \circ(\circ(x+y)+z) \neq \circ(x+\circ(y+z)).$$

- Catastrophic cancellation: 2 floating-point operations are enough to produce a result with relative error $\geqslant 1$.

Catastrophic cancellation

For example, if $x = 1$, $y = \frac{u}{\beta}$, and $z = -1$ then

$$\widehat{r} := \mathrm{RN}(\mathrm{RN}(x + y) + z) = 0,$$

and, since $r := x + y + z$ is nonzero, we obtain $|\widehat{r} - r|/|r| = 1$.

Catastrophic cancellation

For example, if $x = 1$, $y = \frac{u}{\beta}$, and $z = -1$ then

$$\widehat{r} := \mathrm{RN}(\mathrm{RN}(x + y) + z) = 0,$$

and, since $r := x + y + z$ is nonzero, we obtain $|\widehat{r} - r|/|r| = 1$.

Possible workarounds:

- Sorting the input (if possible)
- Rewriting:
$$a^2 - b^2 = (a + b)(a - b).$$
- Compensation: compute the rounding errors, and use them later in the algorithm in order to compensate for their effect.

[Kahan, Rump, ...]

Conditions for exact subtraction

Sterbenz' lemma: [Sterbenz'74]

$$x, y \in \mathbb{F}, \qquad \frac{y}{2} \leqslant x \leqslant 2y \qquad \Rightarrow \qquad x - y \in \mathbb{F}.$$

Conditions for exact subtraction

Sterbenz' lemma: [Sterbenz'74]

$$x, y \in \mathbb{F}, \qquad \frac{y}{2} \leqslant x \leqslant 2y \qquad \Rightarrow \qquad x - y \in \mathbb{F}.$$

- Valid for any base β.
- Applications: Cody and Waite's range reduction, Kahan's accurate algorithms (discriminants, triangle area), ...

Conditions for exact subtraction

Sterbenz' lemma: [Sterbenz'74]

$$x, y \in \mathbb{F}, \qquad \frac{y}{2} \leqslant x \leqslant 2y \qquad \Rightarrow \qquad x - y \in \mathbb{F}.$$

- Valid for any base β.
- Applications: Cody and Waite's range reduction, Kahan's accurate algorithms (discriminants, triangle area), ...
- Proof: [Hauser'96]
  - assume $0 < y \leqslant x \leqslant 2y$.
  - $\mathrm{ulp}(y) \leqslant \mathrm{ulp}(x) \;\Rightarrow\; x - y \in \beta^e\mathbb{Z}$ with $\beta^e = \mathrm{ulp}(y)$.
  - $\frac{x-y}{\beta^e}$ is an integer such that $0 \leqslant \frac{x-y}{\beta^e} \leqslant \frac{y}{\mathrm{ulp}(y)} < \beta^p$. $\square$

Representable error terms

Addition and multiplication:

$$x, y \in \mathbb{F}, \quad \mathsf{op} \in \{+, \times\} \qquad \Rightarrow \qquad x \,\mathsf{op}\, y - \mathsf{RN}(x \,\mathsf{op}\, y) \in \mathbb{F}.$$

Representable error terms

Addition and multiplication:

$$x, y \in \mathbb{F}, \quad \mathrm{op} \in \{+, \times\} \qquad \Rightarrow \qquad x \,\mathrm{op}\, y - \mathrm{RN}(x \,\mathrm{op}\, y) \in \mathbb{F}.$$

Division and square root:

$$x - y\,\mathrm{RN}(x/y) \in \mathbb{F}, \qquad x - \mathrm{RN}\left(\sqrt{x}\right)^2 \in \mathbb{F}.$$

- Noted quite early. [Dekker'71, Pichat'76, Bohlender et al.'91]
- RN required only for ADD and SQRT. [Boldo & Daumas'03]

FMA: its error is the sum of *two* floats. [Boldo & Muller'11]

Error-free transformations (EFT)

Floating-point algorithms for computing such error terms exactly:

- $x + y - \mathrm{RN}(x + y)$ in 6 additions [Møller'65, Knuth] and not less [Kornerup, Lefèvre, Louvet, Muller'12]

Error-free transformations (EFT)

Floating-point algorithms for computing such error terms exactly:

- $x + y - \mathrm{RN}(x + y)$ in 6 additions [Møller'65, Knuth] and not less [Kornerup, Lefèvre, Louvet, Muller'12]
- $xy - \mathrm{RN}(xy)$ can be obtained
  - in 17 + and x [Dekker'71, Boldo'06]
  - in only 2 ops if an FMA is available:

$$\widehat{z} := \mathrm{RN}(xy) \quad \Rightarrow \quad xy - \widehat{z} = \mathrm{FMA}(x, y, -\widehat{z}).$$

Error-free transformations (EFT)

Floating-point algorithms for computing such error terms exactly:

- $x + y - \mathrm{RN}(x + y)$ in 6 additions [Møller'65, Knuth] and not less [Kornerup, Lefèvre, Louvet, Muller'12]
- $xy - \mathrm{RN}(xy)$ can be obtained
  - in 17 + and x [Dekker'71, Boldo'06]
  - in only 2 ops if an FMA is available:

$$\widehat{z} := \mathrm{RN}(xy) \quad \Rightarrow \quad xy - \widehat{z} = \mathrm{FMA}(x, y, -\widehat{z}).$$

- Similar FMA-based EFT for DIV, SQRT ... and FMA.

EFT are key for extended precision algorithms: *error compensation* [Kahan'65, ..., Higham'96, Ogita, Rump, Oishi'04+, Graillat, Langlois, Louvet'05+, ...], *floating-point expansions* [Priest'91, Shewchuk'97, Joldes, Muller, Popescu'14+].

Optimal relative error bounds

When t can be any real number, $E_1(t) \leqslant \frac{u}{1+u}$ and $E_2(t) \leqslant u$ are best possible:

$$t := 1 + u \quad \Rightarrow \quad \mathrm{RN}(t) \text{ is } 1 \text{ or } 1 + 2u \quad \Rightarrow \quad |t - \mathrm{RN}(t)| = u.$$

Optimal relative error bounds

When t can be any real number, $E_1(t) \leqslant \frac{u}{1+u}$ and $E_2(t) \leqslant u$ are best possible:

$$t := 1 + u \quad \Rightarrow \quad \mathrm{RN}(t) \text{ is } 1 \text{ or } 1 + 2u \quad \Rightarrow \quad |t - \mathrm{RN}(t)| = u.$$

Hence

$$E_1(t) = \frac{u}{1+u}$$

Optimal relative error bounds

When t can be any real number, $E_1(t) \leqslant \frac{u}{1+u}$ and $E_2(t) \leqslant u$ are best possible:

$$t := 1 + u \quad \Rightarrow \quad \mathrm{RN}(t) \text{ is } 1 \text{ or } 1 + 2u \quad \Rightarrow \quad |t - \mathrm{RN}(t)| = u.$$

Hence

$$E_1(t) = \frac{u}{1+u}$$

and, if rounding ties "to even", $\mathrm{RN}(t) = 1$ and thus

$$E_2(t) = u.$$

Optimal relative error bounds

When t can be any real number, $E_1(t) \leqslant \frac{u}{1+u}$ and $E_2(t) \leqslant u$ are best possible:

$$t := 1 + u \quad \Rightarrow \quad \mathrm{RN}(t) \text{ is } 1 \text{ or } 1 + 2u \quad \Rightarrow \quad |t - \mathrm{RN}(t)| = u.$$

Hence

$$E_1(t) = \frac{u}{1+u}$$

and, if rounding ties “to even”, $\mathrm{RN}(t) = 1$ and thus

$$E_2(t) = u.$$

These are examples of **optimal bounds**:

- valid for all (t, RN) with t of a certain type;
- attained for some (t, RN) with t parametrized by β and p.

Can we do better when $t = x \,\mathbf{op}\, y$ and $x, y \in \mathbb{F}$?

This depends on *op* and, sometimes, on β and *p*. [J. & Rump'14]

Can we do better when $t = x \,\mathrm{op}\, y$ and $x, y \in \mathbb{F}$?

This depends on *op* and, sometimes, on β and p. [J. & Rump'14]

t	optimal bound on $E_1(t)$	optimal bound on $E_2(t)$
$x \pm y$	$\frac{u}{1+u}$	u
xy	$\frac{u}{1+u}$ $(\star)$	u $(\star)$
x/y	$\begin{cases} \frac{u}{1+u} & \text{if } \beta > 2, \\ u - 2u^2 & \text{if } \beta = 2 \end{cases}$	$\begin{cases} u & \text{if } \beta > 2, \\ \frac{u-2u^2}{1+u-2u^2} & \text{if } \beta = 2 \end{cases}$
$\sqrt{x}$	$1 - \frac{1}{\sqrt{1+2u}}$	$\sqrt{1+2u} - 1$

$(\star)$ iff $\beta > 2$ or $2^p + 1$ is not a Fermat prime.

$\longrightarrow$ **Two standard models for *each* arithmetic operation.**
$\longrightarrow$ **Application: sharper bounds and/or much simpler proofs.**

Context

Floating-point arithmetic

A priori analysis

Conclusion

Classical approach: Wilkinson's analysis

- This is the most common way to guarantee a priori that the computed solution $\widehat{x}$ has some kind of numerical quality:
  - the forward error $\|x - \widehat{x}\|$ is 'small',
  - the backward error $\|\Delta A\|$ such that $(A + \Delta A)\widehat{x} = b$ is 'small'.

Classical approach: Wilkinson's analysis

- This is the most common way to guarantee a priori that the computed solution $\widehat{x}$ has some kind of numerical quality:
  - the forward error $\|x - \widehat{x}\|$ is 'small',
  - the backward error $\|\Delta A\|$ such that $(A + \Delta A)\widehat{x} = b$ is 'small'.
- Developed by Wilkinson in the 1950s and 1960s.
- Relies almost exclusively on the first standard model:

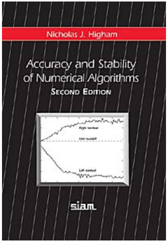
$$\widehat{r} = (x \operatorname{op} y)(1 + \delta), \qquad |\delta| \leqslant u.$$

Classical approach: Wilkinson's analysis

- This is the most common way to guarantee a priori that the computed solution $\widehat{x}$ has some kind of numerical quality:
  - the forward error $\|x - \widehat{x}\|$ is 'small',
  - the backward error $\|\Delta A\|$ such that $(A + \Delta A)\widehat{x} = b$ is 'small'.
- Developed by Wilkinson in the 1950s and 1960s.
- Relies almost exclusively on the first standard model:

$$\widehat{r} = (x \operatorname{op} y)(1 + \delta), \qquad |\delta| \leqslant u.$$

- Eminently powerful:
  see Higham's book
  Accuracy and Stability of Numerical Algorithms (SIAM).



Example of analysis: $a^2 - b^2$

Applying the standard model to each operation gives:

$$\begin{aligned}\widehat{r} &= \big(\mathrm{RN}(a^2) - \mathrm{RN}(b^2)\big)(1+\delta_3)\\ &= \big(a^2(1+\delta_1) - b^2(1+\delta_2)\big)(1+\delta_3), \qquad |\delta_i| \leqslant u_1.\end{aligned}$$

Example of analysis: $a^2 - b^2$

Applying the standard model to each operation gives:

$$\begin{aligned}\widehat{r} &= \big(\mathrm{RN}(a^2) - \mathrm{RN}(b^2)\big)(1+\delta_3) \\ &= \big(a^2(1+\delta_1) - b^2(1+\delta_2)\big)(1+\delta_3), \qquad |\delta_i| \leqslant u_1 .\end{aligned}$$

$$\Rightarrow \qquad \widehat{r} - r = a^2(\delta_1 + \delta_3 + \delta_1\delta_3) - b^2(\delta_2 + \delta_3 + \delta_2\delta_3).$$

Example of analysis: $a^2 - b^2$

Applying the standard model to each operation gives:

$$\begin{aligned}\widehat{r} &= \big(\mathrm{RN}(a^2) - \mathrm{RN}(b^2)\big)(1+\delta_3) \\ &= \big(a^2(1+\delta_1) - b^2(1+\delta_2)\big)(1+\delta_3), \qquad |\delta_i| \leqslant u_1 .\end{aligned}$$

$$\Rightarrow \qquad \widehat{r} - r = a^2(\delta_1 + \delta_3 + \delta_1\delta_3) - b^2(\delta_2 + \delta_3 + \delta_2\delta_3).$$

$$\Rightarrow \qquad |\widehat{r} - r| \leqslant 2u\,(a^2 + b^2).$$

Example of analysis: $a^2 - b^2$

Applying the standard model to each operation gives:

$$\begin{aligned}\widehat{r} &= \big(\mathrm{RN}(a^2) - \mathrm{RN}(b^2)\big)(1+\delta_3)\\ &= \big(a^2(1+\delta_1) - b^2(1+\delta_2)\big)(1+\delta_3), \qquad |\delta_i| \leqslant u_1.\end{aligned}$$

$$\Rightarrow \qquad \widehat{r} - r = a^2(\delta_1 + \delta_3 + \delta_1\delta_3) - b^2(\delta_2 + \delta_3 + \delta_2\delta_3).$$

$$\Rightarrow \qquad |\widehat{r} - r| \leqslant 2u\,(a^2 + b^2).$$

$$\Rightarrow \qquad \frac{|\widehat{r} - r|}{|r|} \leqslant 2u \times C, \quad \textbf{condition number } C := \frac{a^2 + b^2}{|a^2 - b^2|}.$$

Example of analysis: $a^2 - b^2$

Applying the standard model to each operation gives:

$$\begin{aligned}\widehat{r} &= \big(\mathrm{RN}(a^2) - \mathrm{RN}(b^2)\big)(1+\delta_3)\\ &= \big(a^2(1+\delta_1) - b^2(1+\delta_2)\big)(1+\delta_3), \qquad |\delta_i| \leqslant u_1\,.\end{aligned}$$

$$\Rightarrow \qquad \widehat{r} - r = a^2(\delta_1 + \delta_3 + \delta_1\delta_3) - b^2(\delta_2 + \delta_3 + \delta_2\delta_3).$$

$$\Rightarrow \qquad |\widehat{r} - r| \leqslant 2u\,(a^2 + b^2).$$

$$\Rightarrow \qquad \frac{|\widehat{r} - r|}{|r|} \leqslant 2u \times C, \quad \textbf{condition number } C := \frac{a^2 + b^2}{|a^2 - b^2|}.$$

Bound easy to derive and to interpret:

- If $C = O(1)$ then relative error in $O(u)$: **highly accurate!**
- If $C \approx 1/u$ then relative error upper bounded by ≈ 1: **it could be that catastrophic cancellation occurs.**

Example of analysis: $(a+b)(a-b)$

$$\widehat{r} := \mathrm{RN}\Big(\mathrm{RN}(a+b)\cdot\mathrm{RN}(a-b)\Big)$$

$$= (a+b)(a-b)\cdot(1+\delta_1)(1+\delta_2)(1+\delta_3), \qquad |\delta_i| \leqslant u_1.$$

$$\Rightarrow \qquad \frac{|\widehat{r}-r|}{|r|} \leqslant (1+u)^3 - 1 \leqslant 3u.$$

Always highly accurate!

Floating-point summation

Given $x_1, \ldots, x_n \in \mathbb{F}$, evaluate their sum in any order.

Classical analysis [Wilkinson'60]:

- Apply the standard model $n-1$ times.
- Deduce that the computed value $\widehat{s} \in \mathbb{F}$ satisfies

$$\left|\widehat{s} - \sum_{i=1}^{n} x_i\right| \leqslant \alpha \sum_{i=1}^{n} |x_i|, \qquad \alpha = (1+u)^{n-1} - 1.$$

Floating-point summation

Given $x_1, \ldots, x_n \in \mathbb{F}$, evaluate their sum in any order.

Classical analysis [Wilkinson'60]:

- Apply the standard model $n-1$ times.
- Deduce that the computed value $\widehat{s} \in \mathbb{F}$ satisfies

$$\left|\widehat{s} - \sum_{i=1}^{n} x_i\right| \leqslant \alpha \sum_{i=1}^{n} |x_i|, \qquad \alpha = (1+u)^{n-1} - 1.$$

✓ Easy to derive, valid for any order, asymptotically optimal:

$$\frac{\text{error}}{\text{error bound}} \to 1 \text{ as } u \to 0.$$

Floating-point summation

Given $x_1, \ldots, x_n \in \mathbb{F}$, evaluate their sum in any order.

Classical analysis [Wilkinson'60]:

- Apply the standard model $n-1$ times.
- Deduce that the computed value $\widehat{s} \in \mathbb{F}$ satisfies

$$\left|\widehat{s} - \sum_{i=1}^{n} x_i\right| \leqslant \alpha \sum_{i=1}^{n} |x_i|, \qquad \alpha = (1+u)^{n-1} - 1.$$

✓ Easy to derive, valid for any order, asymptotically optimal:

$$\frac{\text{error}}{\text{error bound}} \to 1 \text{ as } u \to 0.$$

✗ But, even with u replaced by $\frac{u}{1+u}$, $\alpha = (n-1)u + O(u^2)$, which hides a constant. So, classically bounded as

$$\alpha \leqslant \gamma_{n-1}, \qquad \gamma_\ell = \frac{\ell u}{1 - \ell u}, \qquad \ell u < 1. \qquad \text{[Higham'96]}$$

A simpler, $O(u^2)$-free bound

Theorem [Rump'12]

For recursive summation, one can take $\alpha = (n-1)u$.

A simpler, $O(u^2)$-free bound

Theorem [Rump'12]

For recursive summation, one can take $\alpha = (n-1)u$.

To prove this,

- Don't use just the (refined) standard model

$$|\mathrm{RN}(x+y) - (x+y)| \leqslant \tfrac{u}{1+u}|x+y|. \qquad (1)$$

A simpler, $O(u^2)$-free bound

Theorem [Rump'12]

For recursive summation, one can take $\alpha = (n-1)u$.

To prove this,

- Don't use just the (refined) standard model

$$|\mathrm{RN}(x+y) - (x+y)| \leqslant \tfrac{u}{1+u}|x+y|. \tag{1}$$

- But combine it with the lower-level property

$$\begin{aligned} |\mathrm{RN}(x+y) - (x+y)| &\leqslant |f - (x+y)|, \qquad \forall f \in \mathbb{F}, \\ &\leqslant \min\{|x|, |y|\}; \end{aligned} \tag{2}$$

A simpler, $O(u^2)$-free bound

Theorem [Rump'12]

For recursive summation, one can take $\alpha = (n-1)u$.

To prove this,

- Don't use just the (refined) standard model
$$|\mathrm{RN}(x+y) - (x+y)| \leqslant \tfrac{u}{1+u}|x+y|. \tag{1}$$
- But combine it with the lower-level property
$$\begin{aligned} |\mathrm{RN}(x+y) - (x+y)| &\leqslant |f - (x+y)|, \qquad \forall f \in \mathbb{F}, \\ &\leqslant \min\{|x|, |y|\}; \end{aligned} \tag{2}$$
- Conclude by induction on n with a clever case-distinction comparing $|x_n|$ to $u \cdot \sum_{i<n} |x_i|$, and using either (1) or (2).

Wilkinson's bounds revisited

Problem	Classical α	New α	Ref.
summation	$(n-1)u + O(u^2)$	$(n-1)u$	[1]

Wilkinson's bounds revisited

Problem	Classical α	New α		Ref.
summation	$(n-1)u + O(u^2)$	$(n-1)u$		[1]
dot prod., mat. mul.	$nu + O(u^2)$	nu		[1]
Euclidean norm	$(\frac{n}{2}+1)u + O(u^2)$	$(\frac{n}{2}+1)u$		[2]
$Tx = b, \quad A = LU$	$nu + O(u^2)$	nu		[2]
$A = R^T R$	$(n+1)u + O(u^2)$	$(n+1)u$		[2]
x^n (recursive, $\beta = 2$)	$(n-1)u + O(u^2)$	$(n-1)u$	$(\star)$	[3]
product $x_1 x_2 \cdots x_n$	$(n-1)u + O(u^2)$	$(n-1)u$	$(\star)$	[4]
poly. eval. (Horner)	$2nu + O(u^2)$	$2nu$	$(\star)$	[4]

$(\star)$ if $n < c \cdot u^{-1/2}$.

[1]: with Rump'13; [2]: with Rump'14; [3]: Graillat, Lefèvre, Muller'14;
[4]: with Bünger and Rump'14.

Kahan's algorithm for $ad - bc$

Kahan's algorithm uses the FMA to evaluate $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$:

$$\boxed{\begin{array}{l} \widehat{w} := \mathrm{RN}(bc); \\ \widehat{f} := \mathrm{RN}(ad - \widehat{w}); \qquad e := \mathrm{RN}(\widehat{w} - bc); \\ \widehat{r} := \mathrm{RN}(\widehat{f} + e); \end{array}}$$

Kahan's algorithm for $ad - bc$

Kahan's algorithm uses the FMA to evaluate $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$:

$$\boxed{\begin{array}{l} \widehat{w} := \mathrm{RN}(bc); \\ \widehat{f} := \mathrm{RN}(ad - \widehat{w}); \quad e := \mathrm{RN}(\widehat{w} - bc); \\ \widehat{r} := \mathrm{RN}(\widehat{f} + e); \end{array}}$$

- The operation $ad - bc$ is not in IEEE 754, but very common:
  - complex arithmetic,
  - discriminant of a quadratic equation,
  - robust orientation predicates using tests like '$ad - bc > \epsilon$?'
- If evaluated naively, $ad - bc$ leads to highly inaccurate results:

$$\frac{|\widehat{f} - r|}{|r|} \text{ can be of the order of } u^{-1} \gg 1.$$

Kahan's algorithm for $ad - bc$

- Analysis in the standard model [Higham'96]:

$$\frac{|\widehat{r} - r|}{|r|} \leqslant 2u\left(1 + \frac{u|bc|}{2|r|}\right).$$

$\Rightarrow$ high relative accuracy **as long as** $u|bc| \not\gg 2|r|$.

Kahan's algorithm for $ad - bc$

- Analysis in the standard model [Higham'96]:

$$\frac{|\widehat{r} - r|}{|r|} \leqslant 2u\left(1 + \frac{u|bc|}{2|r|}\right).$$

$\Rightarrow$ high relative accuracy **as long as** $u|bc| \not\gg 2|r|$.

- When $u|bc| \gg 2|r|$, the error bound can be > 1 and does not even allow to conclude that $\mathrm{sign}(\widehat{r}) = \mathrm{sign}(r)$.

Kahan's algorithm for $ad - bc$

- Analysis in the standard model [Higham'96]:

$$\frac{|\widehat{r} - r|}{|r|} \leqslant 2u\left(1 + \frac{u|bc|}{2|r|}\right).$$

$\Rightarrow$ high relative accuracy **as long as** $u|bc| \not\gg 2|r|$.

- When $u|bc| \gg 2|r|$, the error bound can be > 1 and does not even allow to conclude that $\mathrm{sign}(\widehat{r}) = \mathrm{sign}(r)$.

In fact, Kahan's algorithm is **always** highly accurate:

✗ the standard model alone fails to predict this;

✗ misinterpreting bounds $\Rightarrow$ dismissing good algorithms.

Further analysis

[J., Louvet, Muller'13]

The key is an ulp-analysis of the error terms ϵ_1 and ϵ_2 given by:

$$\boxed{\begin{array}{l} \widehat{w} := \mathrm{RN}(bc); \\ \widehat{f} := \mathrm{RN}(ad - \widehat{w}); \quad e := \mathrm{RN}(\widehat{w} - bc); \\ \widehat{r} := \mathrm{RN}(\widehat{f} + e); \end{array}} \qquad \begin{array}{c} \widehat{f} = ad - \widehat{w} + \epsilon_1 \\ \widehat{r} = \widehat{f} + e + \epsilon_2 \end{array}$$

- Since e is exactly $\widehat{w} - bc$, we have $\widehat{r} - r = \epsilon_1 + \epsilon_2$.
- Furthermore, we can prove that $|\epsilon_i| \leqslant \frac{\beta}{2}\mathrm{ulp}(r)$ for $i = 1, 2$.

Proposition: $|\widehat{r} - r| \leqslant \beta\, \mathrm{ulp}(r) \leqslant 2\beta u\, |r|$.

Further analysis

[J., Louvet, Muller'13]

The key is an ulp-analysis of the error terms ϵ_1 and ϵ_2 given by:

$$\boxed{\begin{array}{ll}\widehat{w} := \mathrm{RN}(bc); & \\ \widehat{f} := \mathrm{RN}(ad - \widehat{w}); & e := \mathrm{RN}(\widehat{w} - bc); \\ \widehat{r} := \mathrm{RN}(\widehat{f} + e); & \end{array}} \qquad \begin{array}{l}\widehat{f} = ad - \widehat{w} + \epsilon_1 \\ \widehat{r} = \widehat{f} + e + \epsilon_2\end{array}$$

- Since e is exactly $\widehat{w} - bc$, we have $\widehat{r} - r = \epsilon_1 + \epsilon_2$.
- Furthermore, we can prove that $|\epsilon_i| \leqslant \frac{\beta}{2}\mathrm{ulp}(r)$ for $i = 1, 2$.

Proposition: $|\widehat{r} - r| \leqslant \beta\, \mathrm{ulp}(r) \leqslant 2\beta u\, |r|$.

These bounds mean Kahan's algorithm is **always** highly accurate.

Further analysis

[J., Louvet, Muller'13]

We can do better via a case analysis comparing $|\epsilon_2|$ to $\frac{1}{2}\mathrm{ulp}(r)$:

Theorem:

- relative error $|\widehat{r} - r|/|r| \leqslant 2u$;

Further analysis [J., Louvet, Muller'13]

We can do better via a case analysis comparing $|\epsilon_2|$ to $\frac{1}{2}\text{ulp}(r)$:

Theorem:

- relative error $|\widehat{r} - r|/|r| \leqslant 2u$;
- the leading constant 2 is best possible.

Certificate of optimality

This is an explicit input set parametrized by β and p such that

$$\frac{\text{error}}{\text{error bound}} \quad \to \quad 1 \quad \text{as} \quad u \to 0.$$

Certificate of optimality

This is an explicit input set parametrized by β and p such that

$$\frac{\text{error}}{\text{error bound}} \quad \to \quad 1 \quad \text{as} \quad u \to 0.$$

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

$$\left.\begin{aligned} a &= b = \beta^{p-1} + 1 \\ c &= \beta^{p-1} + \tfrac{\beta}{2}\beta^{p-2} \\ d &= 2\beta^{p-1} + \tfrac{\beta}{2}\beta^{p-2} \end{aligned}\right\} \quad \Rightarrow \quad \frac{|\widehat{r} - r|/|r|}{2u} = \frac{1}{1+2u} = 1 - 2u + O(u^2).$$

Certificate of optimality

This is an explicit input set parametrized by β and p such that

$$\frac{\text{error}}{\text{error bound}} \quad \to \quad 1 \quad \text{as} \quad u \to 0.$$

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

$$\left.\begin{aligned} a &= b = \beta^{p-1} + 1 \\ c &= \beta^{p-1} + \tfrac{\beta}{2}\beta^{p-2} \\ d &= 2\beta^{p-1} + \tfrac{\beta}{2}\beta^{p-2} \end{aligned}\right\} \quad \Rightarrow \quad \frac{|\widehat{r} - r|/|r|}{2u} = \frac{1}{1+2u} = 1 - 2u + O(u^2).$$

- Optimality is asymptotic, but often OK in practice: for $\beta = 2$ and $p = 11$, the above example has relative error $1.999024\ldots u$.
- The certificate consists of sparse, symbolic floating-point data, which we can handle automatically. [J., Louvet, Muller, Plet]

Context

Floating-point arithmetic

A priori analysis

Conclusion

Summary

Floating-point arithmetic is

- specified rigorously by IEEE 754,
- highly structured and much richer than the standard model.

Exploiting this structure leads to enhanced a priori analysis:

- optimal standard models for basic arithmetic operations,
- simpler and sharper Wilkinson-like bounds,
- proofs of nice behavior of some numerical kernels.

Summary

Floating-point arithmetic is

- specified rigorously by IEEE 754,
- highly structured and much richer than the standard model.

Exploiting this structure leads to enhanced a priori analysis:

- optimal standard models for basic arithmetic operations,
- simpler and sharper Wilkinson-like bounds,
- proofs of nice behavior of some numerical kernels.

On-going research:

- consider directed roundings as well.
- take underflow and overflow into account.