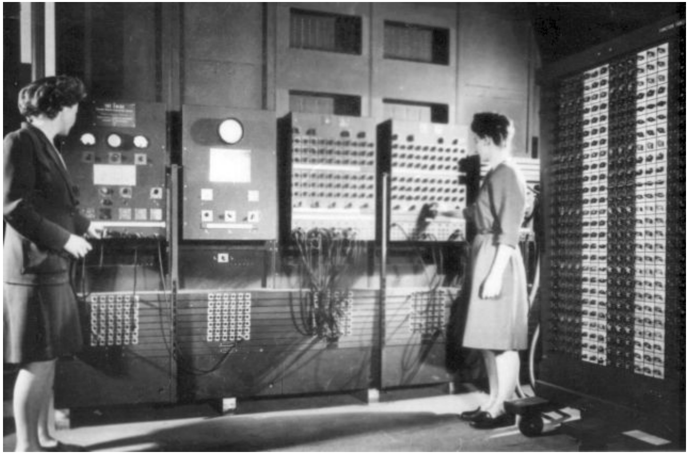
Theoretical and practical aspects of computer arithmetic

Siegfried M. Rump, Hamburg/Tokyo

The origin of floating-point

Computer	base	arithmetic	method	Turing complete
Zuse Z3	binary	floating-point	relais	yes
Atanasoff-Berry	binary	fixed point	tubes	no [linsys $n < 30$]
Colossus	binary	fixed point	tubes	no [deciphering]
Mark I	decimal	fixed point	relais	yes
Eniac	decimal	fixed point	tubes	yes
Babbage	decimal	fixed point	mechanical	yes [not built]



The origin of error analysis I

Carl-Friedrich Gauß was fully aware of computational errors and developed a complete and rigorous error analysis

Based on his computations Ceres was rediscovered

The origin of error analysis II

In their seminal paper

Numerical inverting of matrices of high order (1947)

John v. Neumann and Hermann Goldstine stated:

> "Cholesky decomposition in 24-bit fixed point arithmetic may produce reliable results up to dimension $n \leq 9$."

The analysis is correct but far too pessimistic

Limits of computer arithmetic

Let $\mathbb{A} \subseteq \mathbb{R}$ with $|\mathbb{A}| < \infty$.

There is no isomorphism from $\mathbb{R}$ to $\mathbb{A}$.

There is no meaningful homomorphism respecting order relations.

Back
Close

Limits of computer arithmetic

Let $\mathbb{A} \subseteq \mathbb{R}$ with $|\mathbb{A}| < \infty$.

There is no isomorphism from $\mathbb{R}$ to $\mathbb{A}$.

There is no meaningful homomorphism respecting order relations.

Under very general assumptions it can be shown that operations on $\mathbb{A}$ cannot meet the law of associativity or distributivity.

That is due to the finiteness of $\mathbb{A}$.

The IEEE 754 arithmetic standard 1984 - a closer look

$$\pm 1.m_1 m_2 \ldots m_k \cdot 2^e \qquad \text{binary floating-point}$$

$\mathbb{F}$ set of floating-point numbers

Define a mapping (rounding) $\mathrm{fl} : \mathbb{R} \to \mathbb{F}$

Operations $\tilde{\circ} : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ are defined by
$a \,\tilde{\circ}\, b := \mathrm{fl}(a \circ b)$

$$\begin{tikzcd}
\mathbb{F} \times \mathbb{F} \arrow[r, "\circ"] \arrow[dr, "\tilde{\circ}"'] \arrow[dr, phantom, shift left=3ex, "\curvearrowright"] & \mathbb{R} \arrow[d, "\mathrm{fl}"] \\
& \mathbb{F}
\end{tikzcd}$$

Back
Close

The IEEE 754 arithmetic standard 1984 - a closer look

$$\pm 1.m_1 m_2 \dots m_k \cdot 2^e \qquad \text{binary floating-point}$$

$\mathbb{F}$ set of floating-point numbers

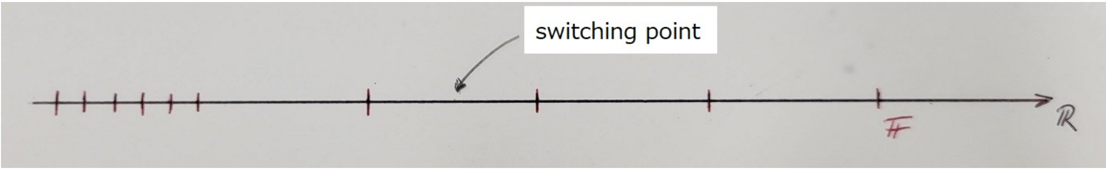
Define a mapping (rounding) $\quad \mathrm{fl} : \mathbb{R} \to \mathbb{F}$

Operations $\tilde{\circ} : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ are defined by $a \,\tilde{\circ}\, b := \mathrm{fl}(a \circ b)$

$$\begin{tikzcd} \mathbb{F} \times \mathbb{F} \arrow[rr, "\circ"] \arrow[drr, "\tilde{\circ}"'] \arrow[rr, phantom, shift right=6ex, "\curvearrowright"] & & \mathbb{R} \arrow[d, "\mathrm{fl}"] \\ & & \mathbb{F} \end{tikzcd}$$

In rounding to nearest, the mapping $\mathrm{fl}_\square$ has minimal error:

$$x \in \mathbb{R} \quad \Rightarrow \quad |\mathrm{fl}_\square(x) - x| = \min\{|f - x| : f \in \mathbb{F}\}$$



The results of arithmetic operations $\tilde{\circ}$ is best possible.

Back
Close

The IEEE 754 arithmetic standard 1984 - a closer look

$$\pm 1.m_1 m_2 \dots m_k \cdot 2^e \qquad \text{binary floating-point}$$

$\mathbb{F}$ set of floating-point numbers

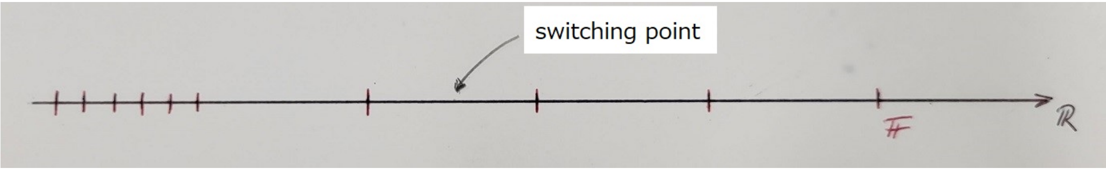
$$\begin{tikzcd} \mathbb{F}\times\mathbb{F} \arrow[r, "\circ"] \arrow[dr, "\tilde{\circ}"'] & \mathbb{R} \arrow[d, "\mathrm{fl}"] \arrow[dl, phantom, "\curvearrowright"] \\ & \mathbb{F} \end{tikzcd}$$

Define a mapping (rounding) $\mathrm{fl} : \mathbb{R} \to \mathbb{F}$

Operations $\tilde{\circ} : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ are defined by
$a \,\tilde{\circ}\, b := \mathrm{fl}(a \circ b)$

In rounding to nearest, the mapping $\mathrm{fl}_\square$ has minimal error:

$$x \in \mathbb{R} \;\Rightarrow\; |\mathrm{fl}_\square(x) - x| = \min\{|f - x| : f \in \mathbb{F}\}$$



The results of arithmetic operations $\tilde{\circ}$ is best possible.

What means "best"?

Back
Close

The relative rounding error - switching points

First standard model $\quad E_1(x) := \left|\frac{\mathrm{fl}(x) - x}{x}\right| \quad$ rel. err. w.r.t. x

Switching point: *arithmetic* mean of adjacent fl-pt numbers

The relative rounding error - switching points

First standard model $\quad E_1(x) := \left|\dfrac{\mathrm{fl}(x) - x}{x}\right| \quad$ rel. err. w.r.t. x

Switching point: *arithmetic* mean of adjacent fl-pt numbers

Second standard model $\quad E_2(x) := \left|\dfrac{\mathrm{fl}(x) - x}{\mathrm{fl}(x)}\right| \quad$ rel. err. w.r.t. $\mathrm{fl}(x)$

Switching point: *harmonic* mean of adjacent fl-pt numbers

The relative rounding error - switching points

First standard model $\quad E_1(x) := \left|\frac{\text{fl}(x) - x}{x}\right| \quad$ rel. err. w.r.t. x

Switching point: *arithmetic* mean of adjacent fl-pt numbers

Second standard model $\quad E_2(x) := \left|\frac{\text{fl}(x) - x}{\text{fl}(x)}\right| \quad$ rel. err. w.r.t. $\text{fl}(x)$

Switching point: *harmonic* mean of adjacent fl-pt numbers

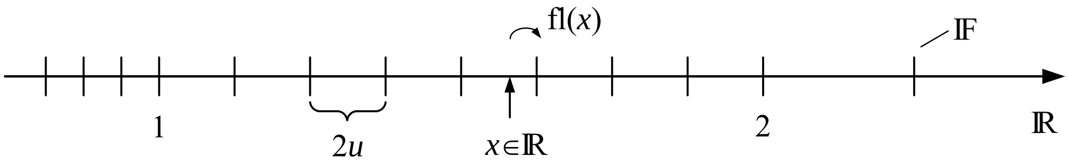
Minimize $\quad \max\{E_1(x), E_2(x)\}$

Switching point: *geometric* mean of adjacent fl-pt numbers

S.M. Rump and M. Lange. On the Definition of Unit Roundoff.
BIT Numerical Mathematics, 56(1):309–317, 2015.

The standard models for the relative rounding error

Rounding to nearest with relative rounding error unit $\mathbf{u}$

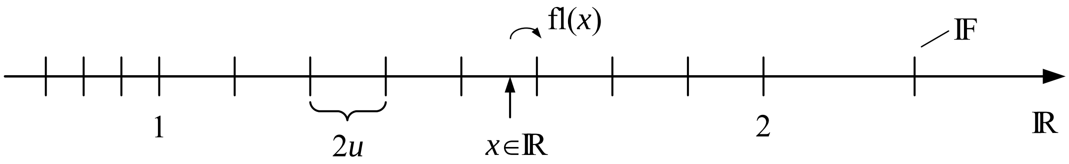


$$x \in [1,2]: \qquad |\mathrm{fl}(x) - x| \leq \mathbf{u}$$

Back
Close

The standard models for the relative rounding error

Rounding to nearest with relative rounding error unit $\mathbf{u}$



$$x \in [1,2]: \qquad |\mathrm{fl}(x) - x| \le \mathbf{u}$$

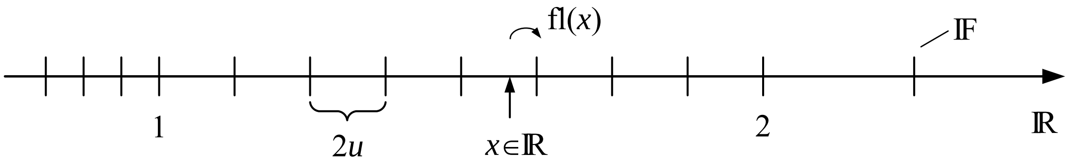
relative rounding error $\quad E_2(x) := \left|\dfrac{\mathrm{fl}(x) - x}{\mathrm{fl}(x)}\right| \le \dfrac{\mathbf{u}}{1} = \mathbf{u} \qquad$ w.r.t. fl(x)

$$\Rightarrow \quad (1+\varepsilon)\mathrm{fl}(x) = x \qquad\qquad |\varepsilon| \le \mathbf{u}$$

Back
Close

The standard models for the relative rounding error

Rounding to nearest with relative rounding error unit $\mathbf{u}$



$$x \in [1,2]: \qquad |\mathrm{fl}(x) - x| \le \mathbf{u}$$

relative rounding error $\quad E_2(x) := \left| \dfrac{\mathrm{fl}(x) - x}{\mathrm{fl}(x)} \right| \le \dfrac{\mathbf{u}}{1} = \mathbf{u} \qquad$ w.r.t. $\mathrm{fl}(x)$

$$\Rightarrow \quad (1+\varepsilon)\mathrm{fl}(x) = x \qquad\qquad |\varepsilon| \le \mathbf{u}$$

$$E_1(x) := \left| \frac{\mathrm{fl}(x) - x}{x} \right| = \left| \frac{\varepsilon\, \mathrm{fl}(x)}{(1+\varepsilon)\mathrm{fl}(x)} \right| = \left| \frac{\varepsilon}{1+\varepsilon} \right| \le \frac{\mathbf{u}}{1+\mathbf{u}} \qquad \text{w.r.t. } x$$

P.H. Sterbenz: Floating-Point Computations, Prentice-Hall, 1974

Back
Close

Optimal bounds of floating-point operations

t	bound on $E_1(t)$	bound on $E_2(t)$
real number	$\frac{\mathbf{u}}{1+\mathbf{u}}$	$\mathbf{u}$
$a \pm b$	$\frac{\mathbf{u}}{1+\mathbf{u}}$	$\mathbf{u}$
ab	$\frac{\mathbf{u}}{1+\mathbf{u}}$	$\mathbf{u}$
a/b	$\begin{cases} \mathbf{u} - 2\mathbf{u}^2 & \text{if } \beta = 2, \\ \frac{\mathbf{u}}{1+\mathbf{u}} & \text{if } \beta > 2 \end{cases}$	$\begin{cases} \frac{\mathbf{u}-2\mathbf{u}^2}{1+\mathbf{u}-2\mathbf{u}^2} & \text{if } \beta = 2, \\ \mathbf{u} & \text{if } \beta > 2 \end{cases}$
$\sqrt{a}$	$1 - \frac{1}{\sqrt{1+2\mathbf{u}}}$	$\sqrt{1+2\mathbf{u}} - 1$

The bounds are optimal for p-digit base-β IEEE-754 arithmetic under some mild conditions.

For example, multiplication in base $\beta = 2$ requires that

$2^p + 1$ is not a Fermat prime.

C.-P. Jeannerod and S.M. Rump. On relative errors of floating-point operations: Optimal bounds and applications. Mathematics of Computation, 87:803–819, 2018.

Composed operations: Classical Wilkinson-type error estimates

Summation $p_1 + p_2 + \ldots + p_n$

recursive summation $\quad \hat{s} := p_1$

$$\hat{s}_i := \hat{s}_{i-1} \mathbin{\tilde{+}} p_i \quad \text{for } i \in \{2, \ldots, n\}$$

Composed operations: Classical Wilkinson-type error estimates

Summation $p_1 + p_2 + \ldots + p_n$

recursive summation $\quad \hat{s} := p_1$

$$\hat{s}_i := \hat{s}_{i-1} \tilde{+} p_i \quad \text{for } i \in \{2, \ldots, n\}$$

... now "Epsilontik" starts

classical $\hat{s}_n = (\ldots((p_1 + p_2)(1+\varepsilon_1) + p_3)(1+\varepsilon_2) + \ldots p_n)(1+\varepsilon_{n-1})$

$$\Rightarrow \quad \left| \hat{s}_n - \sum_{i=1}^{n} p_i \right| \le \left((1+\mathbf{u})^{n-1} - 1\right) \sum_{i=1}^{n} |p_i| \qquad \le \underbrace{\frac{(n-1)\mathbf{u}}{1-(n-1)\mathbf{u}}}_{\gamma_{n-1}} \sum_{i=1}^{n} |p_i|$$

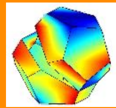
[provided that $(n-1)\mathbf{u} < 1$]

Classical since the 1960's $\qquad$ but not "nice"

Linearized bounds for composed operations !

[R. 2012] $$\left|\hat{s} - \sum_{i=1}^{n} p_i\right| \leq (n-1)\mathbf{u} \sum_{i=1}^{n} |p_i|$$

no limit on n



Linearized bounds for composed operations !

[R. 2012] $\left|\hat{s} - \sum_{i=1}^{n} p_i\right| \le (n-1)\mathbf{u} \sum_{i=1}^{n} |p_i|$ ... the race began

no limit on n

Linearized bounds for composed operations !

[R. 2012] $\left|\hat{s} - \sum_{i=1}^{n} p_i\right| \leq (n-1)\mathbf{u} \sum_{i=1}^{n} |p_i|$ ... the race began

no limit on n

[Jeannerod, R. 2013] $\left|\hat{s} - \sum_{i=1}^{n} x_i\right| \leq n\mathbf{u} \sum_{i=1}^{n} |x_i|$

- $x_i \in \mathbb{R}$
- summation of $\mathrm{fl}(x_i)$ in floating-point
- any base $\beta \geq 2$
- any order of evaluation
- no limit on n

Corollary $|\hat{r} - a^T b| \leq n\mathbf{u}|a^T|\,|b|$ for $a, b \in \mathbb{F}^n$

More linearized bounds for compound operations

[Graillat, Lefèvre, Muller 2015] power

$$|\hat{r} - a^{k+1}| \le k\mathbf{u}|a^{k+1}| \qquad \text{if} \quad k \le \sqrt{2^{1/3} - 1}\,\mathbf{u}^{-1/2} - 1$$

- base $\beta = 2$
- successive multiplication

[R., Bünger, Jeannerod 2015] products

$$\left|\hat{r} - \prod_{i=0}^{k} x_i\right| \le k\mathbf{u}\left|\prod_{i=0}^{k} x_i\right| \qquad \text{for } x_i \in \mathbb{F},\ \beta = 2,\ k < \mathbf{u}^{-1/2}$$

- any order of evaluation
- limit on k is mandatory
- $k < \mathbf{u}^{-1/2}$ cannot be replaced by $k < 12\mathbf{u}^{-1/2}$

More linearized bounds for compound operations (cont'd)

[R., Bünger, Jeannerod 2015] Horner's scheme

$$\left|\hat{r} - \sum_{i=0}^{n} a_i x^i\right| \le 2n\mathbf{u} \sum_{i=0}^{n} |a_i x^i| \qquad \text{if } n < \frac{1}{2}\left(\sqrt{\frac{\omega}{\beta}}\,\mathbf{u}^{-1/2} - 1\right).$$

Classical

$$|\hat{r} - \|p\|_2| \le \left((1+\mathbf{u})^{n/2+1} - 1\right)\|p\|_2 \quad \text{for } p \in \mathbb{F}^n$$

[Jeannerod, R. 2016]

$$|\hat{r} - \|p\|_2| \le \left(\frac{n}{2} + 1\right)\mathbf{u}\|p\|_2$$

- any order of evaluation
- no restriction on n

Linearized bounds for algorithms

Classical $\gamma_k := \frac{k\mathbf{u}}{1-k\mathbf{u}}, \quad k\mathbf{u} < 1$

- $A \in \mathbb{F}^{m\times n}$, computed LU-factors $\hat{L}, \hat{U}$:
  $\hat{L}\hat{U} = A + \Delta A, \qquad |\Delta A| \le \gamma_n \; |\hat{L}|\,|\hat{U}|$
- $A \in \mathbb{F}^{n\times n}$, computed Cholesky factor $\hat{R}$:
  $\hat{R}^T\hat{R} = A + \Delta A, \qquad |\Delta A| \le \quad \gamma_{n+1} \quad |\hat{R}^T|\,|\hat{R}|$
- $T \in \mathbb{F}^{n\times n}$ triangular, $b \in \mathbb{F}^n$, $\hat{x} = T\backslash b$:
  $(T + \Delta T)\hat{x} = b, \qquad |\Delta T| \le \gamma_n \; |T|$

Linearized bounds for algorithms

Improved [R., Jeannerod (2015)]   no limit on n

- $A \in \mathbb{F}^{m\times n}$, computed LU-factors $\hat{L}, \hat{U}$:
  $\hat{L}\hat{U} = A + \Delta A, \qquad |\Delta A| \le n\mathbf{u}|\hat{L}|\,|\hat{U}|$

- $A \in \mathbb{F}^{n\times n}$, computed Cholesky factor $\hat{R}$:
  $\hat{R}^T\hat{R} = A + \Delta A, \qquad |\Delta A| \le (n+1)\mathbf{u}|\hat{R}^T|\,|\hat{R}|$

- $T \in \mathbb{F}^{n\times n}$ triangular, $b \in \mathbb{F}^n$, $\hat{x} = T\backslash b$:
  $(T + \Delta T)\hat{x} = b, \qquad |\Delta T| \le n\mathbf{u}|T|$

Towards a more general perspective

Up to now:

- We *actively* assumed base-β IEEE-754 conform arithmetic.
- Every result relied on that specific arithmetic.

Next:

- *Passively* identify sufficient assumptions to prove linearized bounds.

$\rightarrowtail$ Understand "Machine numbers" $\mathbb{M}$ as a subset of $\mathbb{R}$

An arithmetic on a general subset of $\mathbb{R}$

$\mathbb{M} \subseteq \mathbb{R}$, $\quad \boxed{\circ} : \mathbb{M} \times \mathbb{M} \to \mathbb{M}$ for $\circ \in \{+, -, \times, /\}$, also $\sqrt{\cdot}$

$$x, y \in \mathbb{M}: \quad x \boxed{\circ} y = (x \circ y)(1 + \delta) \qquad |\delta| \le eps$$

for some constant eps. We do *not* assume a rounding function fl !

An arithmetic on a general subset of $\mathbb{R}$

$\mathbb{M} \subseteq \mathbb{R}, \quad \boxed{\circ} : \mathbb{M} \times \mathbb{M} \to \mathbb{M} \quad$ for $\circ \in \{+, -, \times, /\}$, also $\sqrt{\cdot}$

$$x, y \in \mathbb{M}: \quad x \boxed{\circ} y = (x \circ y)(1 + \delta) \qquad |\delta| \le eps$$

for some constant *eps*. We do *not* assume a rounding function fl !

Much freedom:

- $x \circ y \in \mathbb{M} \quad \not\Rightarrow \quad x \boxed{\circ} y = x \circ y$
- $a \circ b = c \circ d \quad \not\Rightarrow \quad a \boxed{\circ} b = c \boxed{\circ} d$

Example 3-digit decimal format, $p = 3$, $eps = \frac{1}{2}\beta^{1-p} = 0.005$

$$x + y = 9.96$$

$$\Rightarrow \; x \boxed{+} y \in \{ 9.92, 9.93, 9.94, 9.95, 9.96, 9.97, 9.98, 9.99, 10.0 \}$$

Back
Close

An arithmetic on a general subset of $\mathbb{R}$

$\mathbb{M} \subseteq \mathbb{R}, \quad \boxed{\circ} : \mathbb{M} \times \mathbb{M} \to \mathbb{M}$ for $\circ \in \{+, -, \times, /\}$, also $\sqrt{\cdot}$

$$x, y \in \mathbb{M}: \quad x \boxed{\circ} y = (x \circ y)(1 + \delta) \qquad |\delta| \le eps$$

for some constant eps. We do *not* assume a rounding function fl !

Much freedom:

- $x \circ y \in \mathbb{M} \quad \not\Rightarrow \quad x \boxed{\circ} y = x \circ y$
- $a \circ b = c \circ d \quad \not\Rightarrow \quad a \boxed{\circ} b = c \boxed{\circ} d$

Example 3-digit decimal format, $p = 3$, $eps = \frac{1}{2}\beta^{1-p} = 0.005$

$$x + y = 9.96$$

$$\Rightarrow \; x \boxed{+} y \in \{9.92, 9.93, 9.94, 9.95, 9.96, 9.97, 9.98, 9.99, 10.0\}$$

e.g. $9.90 \boxed{+} 0.06 = 10 \qquad 9.91 \boxed{+} 0.05 = 9.92$

Back
Close

An arithmetic on a general subset of $\mathbb{R}$

$\mathbb{M} \subseteq \mathbb{R}$, $\quad \boxed{\circ} : \mathbb{M} \times \mathbb{M} \to \mathbb{M}$ for $\circ \in \{+, -, \times, /\}$, also $\sqrt{\cdot}$

$$x, y \in \mathbb{M} : \quad x \boxed{\circ} y = (x \circ y)(1 + \delta) \qquad |\delta| \le eps$$

for some constant eps. We do *not* assume a rounding function fl !

Much freedom:

- $x \circ y \in \mathbb{M} \quad \not\Rightarrow \quad x \boxed{\circ} y = x \circ y$ also $x \boxed{\circ} y$ may change
- $a \circ b = c \circ d \quad \not\Rightarrow \quad a \boxed{\circ} b = c \boxed{\circ} d$

Example 3-digit decimal format, $p = 3$, $\mathbf{u} = \frac{1}{2}\beta^{1-p} = 0.005$

$$x + y = 9.96$$

$$\Rightarrow \; x \boxed{+} y \in \{ 9.92, 9.93, 9.94, 9.95, 9.96, 9.97, 9.98, 9.99, 10.0 \}$$

e.g. $\quad 9.90 \boxed{+} 0.06 = 10 \qquad 9.91 \boxed{+} 0.05 = 9.92 \qquad 9.91 \boxed{+} 0.05 = 9.96$

Linearized bounds: An even simplified exposition

$\forall a, b \in \mathbb{M}$: $\quad \left| (a \boxed{+} b) - (a + b) \right| \leq \min(|a|, |b|)$ Assumption A

Linearized bounds: An even simplified exposition

$\forall a, b \in \mathbb{M}$: $\quad |(a \boxed{+} b) - (a + b)| \le \min(|a|, |b|)$ Assumption A

Very weak: $\quad |3 \boxed{+} 4 - (3 + 4)| \le \min(3, 4) = 3$

Linearized bounds: An even simplified exposition

$\forall\, a, b \in \mathbb{M}$: $\quad |\,(a \boxed{+} b) - (a+b)\,| \le \min(|a|, |b|)$ $\quad$ Assumption A

Very weak: $\quad |3 \boxed{+} 4 - (3+4)| \le \min(3,4) = 3$

IEEE-754 $\quad x \in \mathbb{R}$: $\quad |\mathrm{fl}(x) - x| = \min\{|f - x| : f \in \mathbb{F}\}$ $\quad$ nearest

$$
\begin{aligned}
\Rightarrow \quad |a \boxed{+} b - (a+b)| &= |\mathrm{fl}(a+b) - (a+b)| \\
&= \min(|f - (a+b)| : f \in \mathbb{F}) \\
&\le \min(|a - (a+b)|, |b - (a+b)|) \\
&= \min(|a|, |b|)
\end{aligned}
$$

Linearized bounds: An even simplified exposition

$$\forall a, b \in \mathbb{M}: \qquad |\,(a \boxed{+} b) - (a+b)\,| \;\le\; \min(|a|, |b|)$$ Assumption A

$$\text{Very weak:} \qquad |3 \boxed{+} 4 - (3+4)| \;\le\; \min(3,4) = 3$$

IEEE-754 $\quad x \in \mathbb{R}: \quad |\mathrm{fl}(x) - x| \;=\; \min\{|f - x| : f \in \mathbb{F}\}$ $\quad$ nearest

$$\begin{aligned}
\Rightarrow \qquad |a \boxed{+} b - (a+b)| \;&=\; |\mathrm{fl}(a+b) - (a+b)| \\
&=\; \min(|f - (a+b)| : f \in \mathbb{F}) \\
&\le\; \min(|a - (a+b)|, |b - (a+b)|) \\
&=\; \min(|a|, |b|)
\end{aligned}$$

Not satisfied for rounding upwards:

$$1 \boxed{+} \mathbf{u}^2 = \mathrm{succ}(1) = 1 + 2\mathbf{u} \quad \Rightarrow \quad 2\mathbf{u} - \mathbf{u}^2 \;\not\le\; \min(1, \mathbf{u}^2) = \mathbf{u}^2$$

The linearized error estimate

Theorem. Let an arithmetic on $\mathbb{M}$ with Assumption A be given. For $p \in \mathbb{M}^n$ define

$$\hat{s}_1 := p_1; \quad \hat{s}_k = \hat{s}_{k-1} \boxed{+} p_k = (\hat{s}_{k-1} + p_k)(1 + \delta_k) \quad \text{for } 2 \le k \le n$$

with $|\delta_k| \le eps$.

Then

$$\left| \hat{s}_n - \sum_{i=1}^{n} p_i \right| \le \sum_{i=1}^{n} |\delta_i| \sum_{i=1}^{n} |p_i| \le (n-1) eps \sum_{i=1}^{n} |p_i| \tag{$*$}$$

The result is true under much more general assumptions

E.g. (*) is true for directed rounding (not satisfying Assumption A)

M. Lange and S.M. Rump. Error estimates for the summation of real numbers with application to floating-point summation. BIT, 57:927–941, 2017.

Optimal bounds for summation

Worst case $1 + \mathbf{u} + \mathbf{u} + \ldots$?

Mascarenhas 2016:

$$\beta = 2, \quad p \in \mathbb{F}^n, \quad n \le \frac{1}{5}\,2^{p-2} : \quad \left| \hat{s} - \sum_{i=1}^{n} p_i \right| \le \frac{(n-1)\mathbf{u}}{1+(n-1)\mathbf{u}} \sum_{i=1}^{n} |p_i|$$

Proof uses some optimization and continuous mathematics

Optimal bounds for summation

Worst case $1 + \mathbf{u} + \mathbf{u} + \ldots$?

Mascarenhas 2016:

$$\beta = 2, \quad p \in \mathbb{F}^n, \quad n \le \frac{1}{5} 2^{p-2} : \qquad \left| \hat{s} - \sum_{i=1}^{n} p_i \right| \le \frac{(n-1)\mathbf{u}}{1+(n-1)\mathbf{u}} \sum_{i=1}^{n} |p_i|$$

Proof uses some optimization and continuous mathematics

<u>Theorem</u> For an arithmetic on $\mathbb{M}$ with Assumption A and $x \in \mathbb{M}^n$

$$\left| \hat{s} - \sum_{i=1}^{n} x_i \right| \le \frac{\sum_{i=1}^{n-1} \xi_i}{1 + \sum_{i=1}^{n-1} \xi_i} \sum_{i=1}^{n} |x_i| \qquad \text{[IEEE-754:} \quad |\xi_i| \le \mathbf{u}]$$

The estimate is sharp.

M. Lange and S.M. Rump. Sharp estimates for perturbation errors in summations.
Math. of Comp., 88:349–368, 2019.

Error-free transformations

```
function [x,y] = TwoSum(a,b)
  x = a + b;
  z = x - a;
  y = ( a - (x-z) ) + (b-z);
```

Knuth 1969: $\quad a, b \in \mathbb{F} \quad \Rightarrow \quad x + y = a + b$

Error-free transformations

```
function [x,y] = TwoSum(a,b)
  x = a + b;
  z = x - a;
  y = ( a - (x-z) ) + (b-z);
```

Knuth 1969: $a, b \in \mathbb{F} \quad \Rightarrow \quad x + y = a + b$

```
function [x,y] = FastTwoSum(a,b)
  x = a + b;
  y = a - (x - b);
```

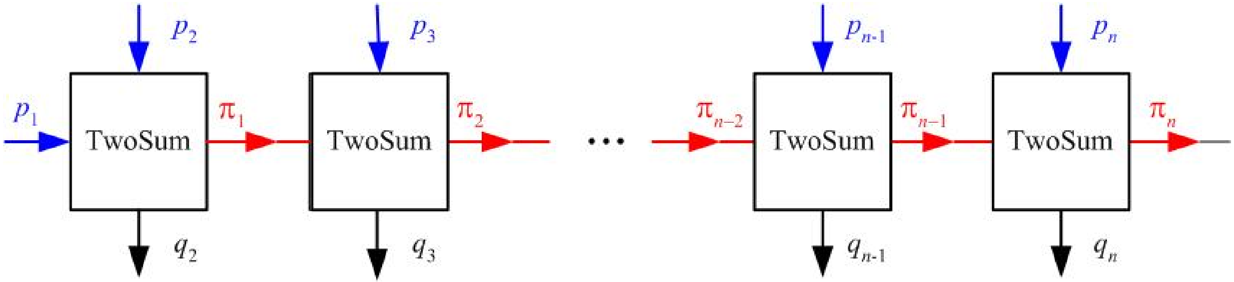
Dekker 1971: $a, b \in \mathbb{F},\ |a| \geq |b| \quad \Rightarrow \quad x + y = a + b$

`FastTwoSum` with comparison often 2 times slower than `TwoSum`

Error-free vector transformations

```
function p = VecSum(p)
  for i=2:n
    [p(i),p(i-1)] = TwoSum(p(i),p(i-1))
```

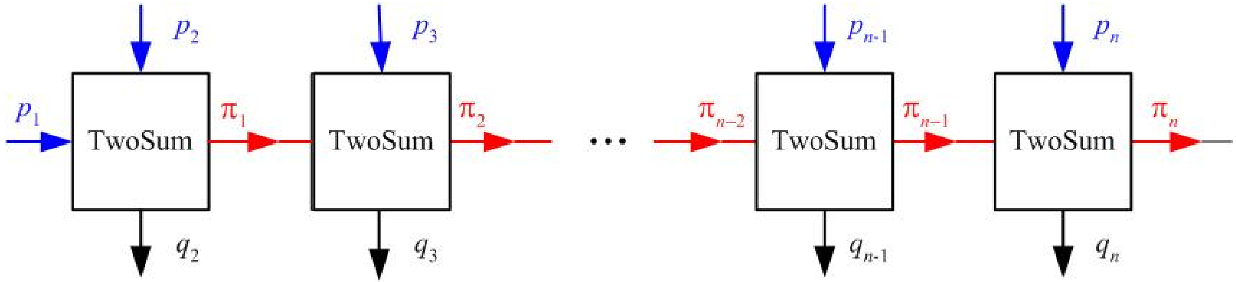
$$q = \mathtt{VecSum}(p) \quad \Rightarrow \quad \sum q_i = \sum p_i, \quad q_n = \mathrm{float}(\sum p_i)$$



Error-free vector transformations

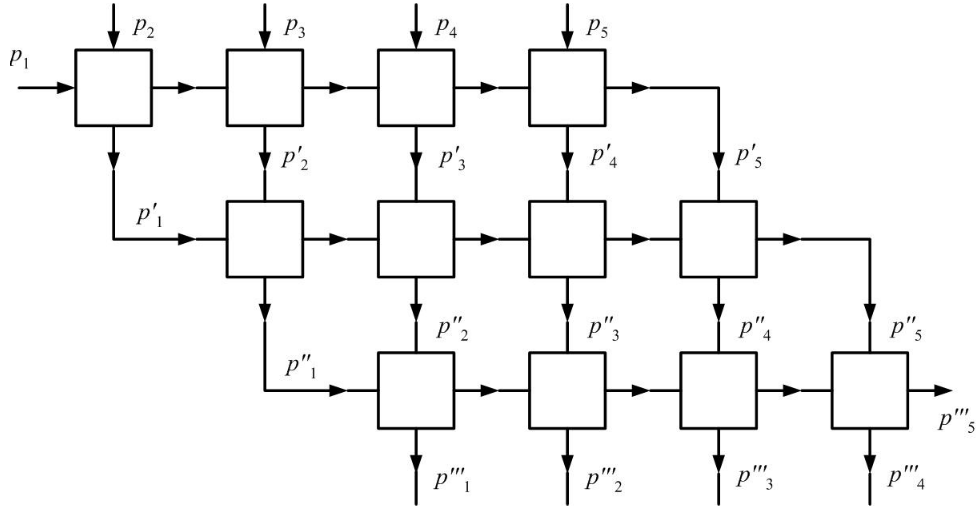
```
function p = VecSum(p)
  for i=2:n
    [p(i),p(i-1)] = TwoSum(p(i),p(i-1))
```

$$q = \mathtt{VecSum}(p) \quad \Rightarrow \quad \sum q_i = \sum p_i, \quad q_n = \mathrm{float}(\sum p_i)$$

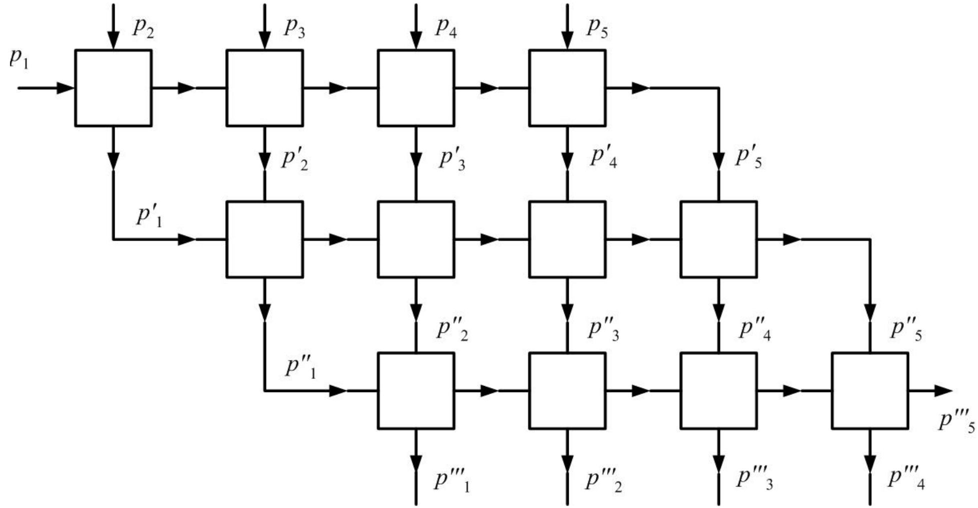


Error of **sum(p)** of the order $[(n-1)\mathbf{u}]^2$

Iterated error-free vector transformations



Iterated error-free vector transformations



Error of `sum(p)` of the order $[(n-1)\mathbf{u}]^{K+1}$ after K transformations

Similar routines for dot products, most important in numerical analysis

T. Ogita, S.M. Rump, and S. Oishi. Accurate sum and dot product. SIAM Journal on Scientific Computing (SISC), 26(6):1955–1988, 2005.

The power of modern error analysis

John v. Neumann and Hermann Goldstine stated:

> "Cholesky decomposition in 24-bit fixed point arithmetic may produce reliable results up to dimension $n \le 9$."

Theorem. Let $A \in \mathbb{F}^{n \times n}$ with $A^T = A$ be given, and let $B = A - D \in \mathbb{F}^{n \times n}$ for diagonal D with $D \ge 2\alpha I$ and $\alpha \ge \gamma_{n+1}\mathrm{trace}(A) > 0$.

The power of modern error analysis

John v. Neumann and Hermann Goldstine stated:

> "Cholesky decomposition in 24-bit fixed point arithmetic may produce reliable results up to dimension $n \le 9$."

Theorem. Let $A \in \mathbb{F}^{n\times n}$ with $A^T = A$ be given, and let $B = A - D \in \mathbb{F}^{n\times n}$ for diagonal D with $D \ge 2\alpha I$ and $\alpha \ge \gamma_{n+1}\mathrm{trace}(A) > 0$.

If the *floating-point* Cholesky decomposition of B runs to completion, then A is symmetric positive definite, and for any $\tilde{x} \in \mathbb{R}^n$

$$\|A^{-1}b - \tilde{x}\|_2 \le \alpha^{-1}\|A\tilde{x} - b\|_2 \, .$$

The power of modern error analysis

John v. Neumann and Hermann Goldstine stated:

> "Cholesky decomposition in 24-bit fixed point arithmetic may produce reliable results up to dimension $n \le 9$."

Theorem. Let $A \in \mathbb{F}^{n\times n}$ with $A^T = A$ be given, and let $B = A - D \in \mathbb{F}^{n\times n}$ for diagonal D with $D \ge 2\alpha I$ and $\alpha \ge \gamma_{n+1}\mathrm{trace}(A) > 0$.

If the *floating-point* Cholesky decomposition of B runs to completion, then A is symmetric positive definite, and for any $\tilde{x} \in \mathbb{R}^n$

$$\|A^{-1}b - \tilde{x}\|_2 \le \alpha^{-1}\|A\tilde{x} - b\|_2 \,.$$

That approach works for dimensions n in the 10-thousands.
All operations are in ordinary floating-point arithmetic !

The analysis is based on properties of a symm. pos. def. matrix

S.M. Rump and T. Ogita. Super-fast validated solution of linear systems. JCAM, 199(2):199–206, 2006.

Towards solving general problems

What about general linear systems, nonlinear systems, global optimization, differential equations etc. ?

We may use interval arithmetic:

$$[a, b] \circ [c, d] := [\min x, \max x] \quad \text{for} \quad x \in \{a \circ c, a \circ d, b \circ c, b \circ d\}$$

On the computer we use directed roundings.

Towards solving general problems

What about general linear systems, nonlinear systems, global optimization, differential equations etc. ?

We may use interval arithmetic:

$$[a,b] \circ [c,d] := [\min x, \max x] \quad \text{for} \quad x \in \{a \circ c, a \circ d, b \circ c, b \circ d\}$$

On the computer we use directed roundings.

Fundamental inclusion property:

$$\forall a \in A, b \in B : \quad a \circ b \in A \circ B \qquad \text{for interval quantities } A, B$$

Covers all elementary standard functions, erf, $\Gamma(x)$ etc. as well

Towards solving general problems

Fundamental observation:

Replace in an algorithm all operations
by the corresponding interval operations.

If finished successfully, i.e., no division by a zero interval, then

- It is mathematically certain that the problem is solvable, and
- the computed results do contain the true solution.

Towards solving general problems

Fundamental observation:

Replace in an algorithm all operations
by the corresponding interval operations.

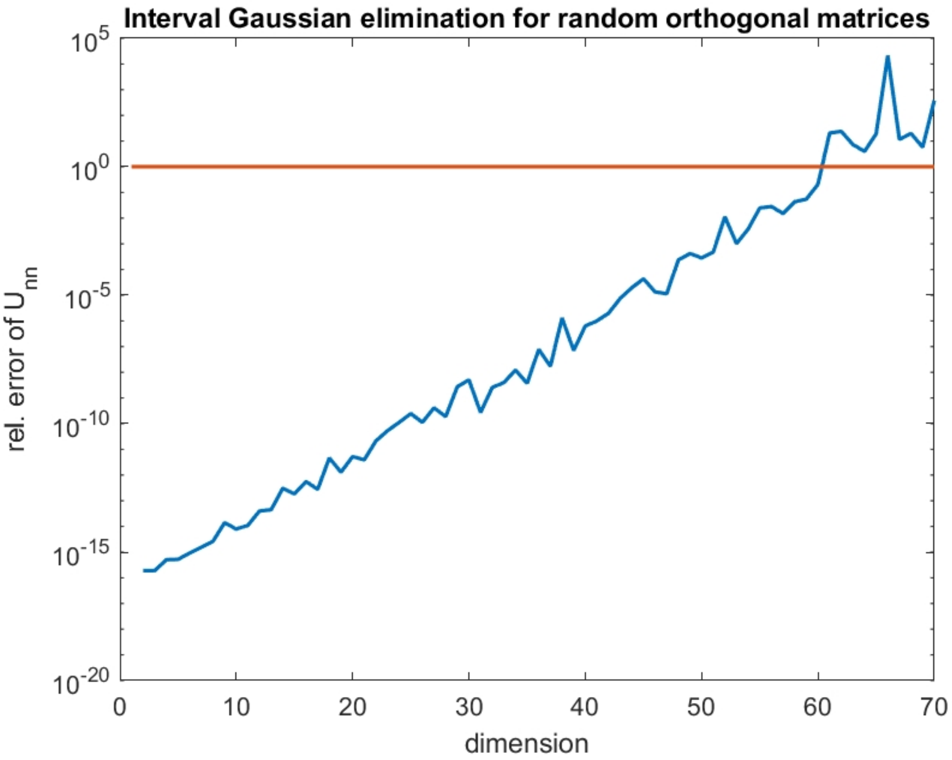
If finished successfully, i.e., no division by a zero interval, then

- It is mathematically certain that the problem is solvable, and
- the computed results do contain the true solution.

This is called *naive* interval arithmetic

Why does interval arithmetic has a bad reputation?

Naive interval arithmetic: Interval Gaussian elimination (IGA)



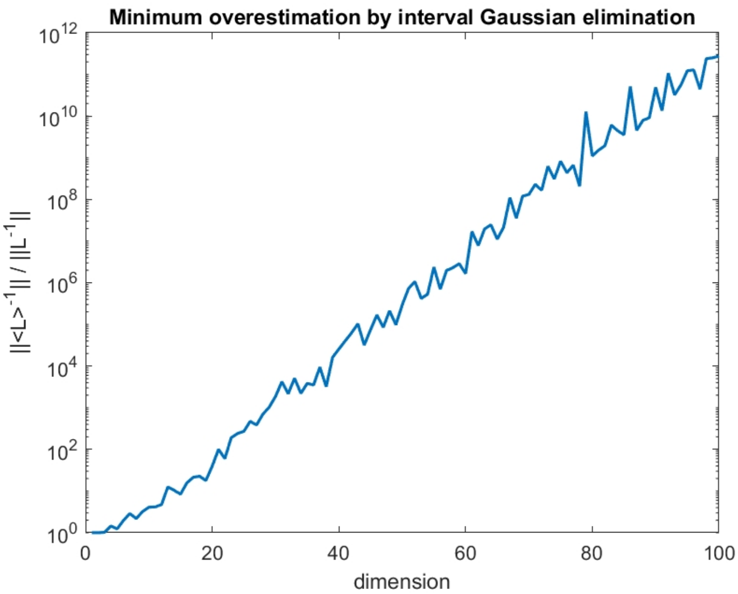
The matrices are perfectly well conditioned: $\quad \operatorname{cond}(A) = 1$

Minimum overestimation for Interval Gaussian elimination (IGA)

Theorem [R., 2010] For $A \in \mathbb{R}^{n\times n}$ perform Gaussian elimination with total pivoting using *real* interval operations everywhere.

If finished successfully, then elementwise

$$\operatorname{rad}(U) \geq \text{upper triangle}\left(< L >^{-1} \cdot \operatorname{rad}(A)\right)$$



The reason for the poor reputation of interval arithmetic

Historically, interval arithmetic was (at least) known to Gauss.

It was tought in German junior high schools from the mid 19th century.

The reason for the poor reputation of interval arithmetic

Historically, interval arithmetic was (at least) known to Gauss.

It was tought in German junior high schools from the mid 19th century.

It was *re-discovered* in the 1960's and *advocated as the holy grail* .

Back
Close

The reason for the poor reputation of interval arithmetic

Historically, interval arithmetic was (at least) known to Gauss.

It was tought in German junior high schools from the mid 19th century.

It was *re-discovered* in the 1960's and *advocated as the holy grail* .

The problem is not the tool [interval arithmetic],

but the way it was used :

Back
Close

Tools — should be used appropriately I

Tools — should be used appropriately II

Is interval arithmetic of any use?

The (unique) advantage of interval arithmetic is to compute bounds for the range of a function over some domain.

The bounds may overestimate the true range, but they are always mathematically true.

A Matlab example ...

How to fight overestimation of interval arithmetic

A verification method should:

- use floating-point arithmetic wherever possible
- try to avoid the dependency problem
- try to scale intervals by a small number

How to fight overestimation of interval arithmetic

A verification method should:

- use floating-point arithmetic wherever possible
- try to avoid the dependency problem
- try to scale intervals by a small number

THEOREM. Let $A, R \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$. If for given $X \in \mathbb{IR}^n$

$$Rb + (I - RA)X \subseteq \mathrm{int}(X)$$

then A is nonsingular and $A^{-1}b \in X$.

How to fight overestimation of interval arithmetic

A verification method should:

- use floating-point arithmetic wherever possible
- try to avoid the dependency problem
- try to scale intervals by a small number

THEOREM. Let $A, R \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$. If for given $X \in \mathbb{IR}^n$

$$Rb + (I - RA)X \subseteq \mathrm{int}(X)$$

then A is nonsingular and $A^{-1}b \in X$.

PROOF. Define $f(x) := Rb + (I - RA)x$. Then

$$\forall x \in X : \quad f(x) \in X \quad \Rightarrow \quad \exists \hat{x} \in X : f(\hat{x}) = \hat{x} = Rb + \hat{x} - RA\hat{x}$$

by Brouwer's fixed point Theorem.

Inclusion in $\mathrm{int}(X)$ implies R, A to be non-singular.

S.M. Rump. Kleine Fehlerschranken bei Matrixproblemen. PhD thesis, Univ. Karlsruhe, 1980.

How to fight overestimation of interval arithmetic

A verification method should:

- use floating-point arithmetic wherever possible
- try to avoid the dependency problem
- try to scale intervals by a small number

THEOREM. Let $A, R \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$. If for given $X \in \mathbb{IR}^n$

$Rb + (I - RA)X \subseteq \mathrm{int}(X)$ do NOT use $X + R(b - AX)$

then A is nonsingular and $A^{-1}b \in X$.

PROOF. Define $f(x) := Rb + (I - RA)x$. Then

$$\forall x \in X : \quad f(x) \in X \quad \Rightarrow \quad \exists \hat{x} \in X : f(\hat{x}) = \hat{x} = Rb + \hat{x} - RA\hat{x}$$

by Brouwer's fixed point Theorem.

Inclusion in $\mathrm{int}(X)$ implies R, A to be non-singular.

S.M. Rump. Kleine Fehlerschranken bei Matrixproblemen. PhD thesis, Univ. Karlsruhe, 1980.

A verification method for systems of nonlinear equations

Let $f : \mathbb{R}^n \to \mathbb{R}^n$, $f \in C^1$, $R \in \mathbb{R}^{n\times n}$, $\tilde{x} \in \mathbb{R}^n$, $X \in \mathbb{IR}^n$. If $\tilde{x} \in X$ and

$$(*) \qquad -Rf(\tilde{x}) + (I - RJ_f(X))X \subseteq \operatorname{int}(X),$$

then there exists a unique root $\hat{x}$ of $f(x) = 0$ in $\tilde{x} + X$.

Verify (*) using interval arithmetic and algorithmic differentiation.

A verification method for systems of nonlinear equations

Let $f : \mathbb{R}^n \to \mathbb{R}^n$, $f \in \mathcal{C}^1$, $R \in \mathbb{R}^{n\times n}$, $\tilde{x} \in \mathbb{R}^n$, $X \in \mathbb{IR}^n$. If $\tilde{x} \in X$ and

$$(*) \qquad -Rf(\tilde{x}) + (I - RJ_f(X))X \subseteq \mathrm{int}(X),$$

then there exists a unique root $\hat{x}$ of $f(x) = 0$ in $\tilde{x} + X$.

Verify (*) using interval arithmetic and algorithmic differentiation.

Rationale, i.e., why is it working well:

$f(\tilde{x}) \approx 0$, $R \approx \frac{\partial f}{\partial x}(\tilde{x})^{-1}$ ensured by good fl-pt approximations

The error w.r.t. to the approximate solution $\tilde{x}$ is included

The product $(I - RJ_f(X))X$ is small in magnitude.

There is a dichotomy:

Either mathematically rigorous inclusion of the solution

or no result (error message)

S.M. Rump. Solving Algebraic Problems with High Accuracy. Habilitation, Acad. Press 1983.

Global optimization in n dimensions

Given $f : \mathbb{R}^n \to \mathbb{R}$, minimize $f(x)$ over a box,
possibly subject to constraints

The main problem: To discard sub-boxes.

Global optimization in n dimensions

Given $f : \mathbb{R}^n \to \mathbb{R}$, minimize $f(x)$ over a box,
possibly subject to constraints

The main problem: To discard sub-boxes.

This is basically outside the scope of (purely) numerical algorithms.

Even if Lipschitz constants are known, rounding errors may have disastrous effects.

Global minimization — Exclusion regions I

(1) Necessarily $\dfrac{\partial f}{\partial x}(\hat{x}) = 0 \quad \to$

If $\quad 0 \notin \left[\dfrac{\partial f}{\partial x}(Y)\right]_i$ for some $\begin{array}{l} 1 \leq i \leq n \\ \text{and} \quad Y \subseteq \operatorname{int}(X) \end{array}$

then $\quad Y$ can be discarded.

Global minimization — Exclusion regions I

(1) Necessarily $\dfrac{\partial f}{\partial x}(\hat{x}) = 0 \quad \rightarrow$

If $\quad 0 \notin \left[\dfrac{\partial f}{\partial x}(Y)\right]_i$ for some $\begin{array}{l} 1 \le i \le n \\ \text{and} \quad Y \subseteq \operatorname{int}(X) \end{array}$

then $\;Y\;$ can be discarded.

(2) Dimension reduction

If $\quad 0 \notin \left[\dfrac{\partial f}{\partial x}(Y)\right]_i$ but $Y_i \bigcap \partial X_i \neq 0$

then $\;Y_i\;$ can be replaced by corresponding ∂X_i.

S.M. Rump. Mathematically Rigorous Global Optimization in Floating-Point Arithmetic. Optimization Methods & Software, 33(4–6):771–798, 2018.

Exclusion regions II — The expansion principle (Jansson)

Let $f : \mathbb{R}^n \to \mathbb{R}, \quad f \in C^1$ be given.

For a given box X, our verification methods can prove that there is exactly one stationary point of f in X.

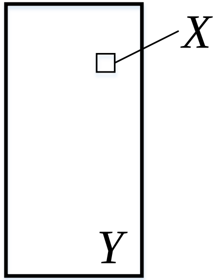
Exclusion regions II — The expansion principle (Jansson)

Let $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^1$ be given.

For a given box X, our verification methods can prove that there is exactly one stationary point of f in X.

Intentionally widen X into $Y \supseteq X$ and suppose that Y as well contains exactly one stationary point.

Then f has no minimum in $Y \backslash X$

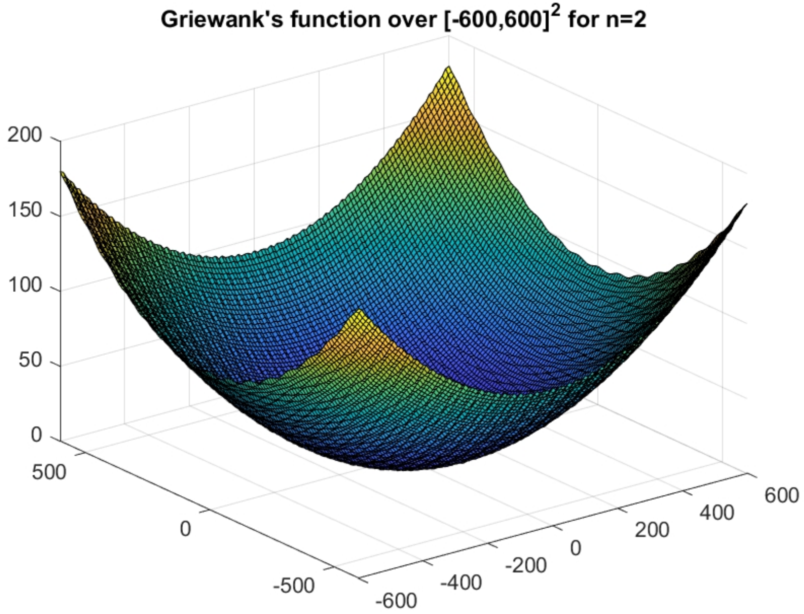


C. Jansson. On Self-Validating Methods for Optimization Problems. In J. Herzberger (ed.) Topics in Validated Computations - Studies in Computational Mathematics 5, 381–438, North-Holland, Amsterdam, 1994.

A famous test function in the global optimization community

Minimize Griewank's function $G : \mathbb{R}^n \to \mathbb{R}$ on $X = [-600, 600]^n$

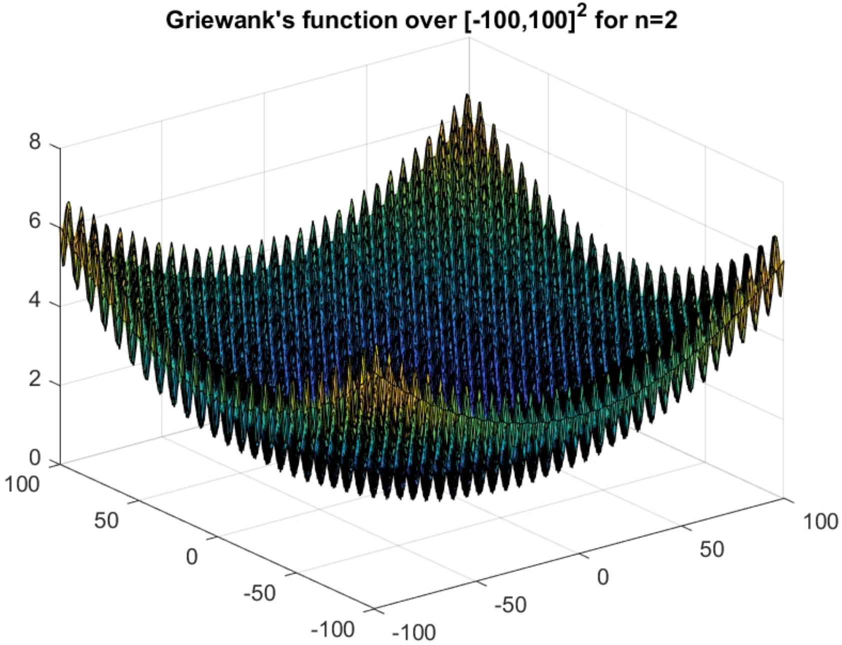
$$G(x) = 1 + \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right)$$



A famous test function

Minimize Griewank's function $G: \mathbb{R}^n \to \mathbb{R}$ on $X = [-600, 600]^n$

$$G(x) = 1 + \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right)$$



A famous test function

Minimize Griewank's function $G: \mathbb{R}^n \to \mathbb{R}$ on $X = [-600, 600]^n$

$$G(x) = 1 + \frac{1}{4000} \sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right)$$

Timing [sec]

n	$\#\nabla G(x) = 0$	Montanher's intsolver	Csendes' GOP	INTLAB
5	$\sim 10^{13}$	307*)	229	0.6
10	$\sim 10^{25}$			1.7
20	$\sim 10^{51}$			5.2
30	$\sim 10^{77}$			10.5
40	$\sim 10^{103}$			17.9
50	$\sim 10^{129}$			28.1

*) verification failed

INTLAB - the Matlab/Octave toolbox for Reliable Computing

- developing since 1998, >2000 routines, >70kLOC pure Matlab
- rigorous input and output
- Real and complex interval arithmetic and standard functions
- affine and Taylor arithmetic
- dense and sparse linear systems
- systems of nonlinear equations
- global optimization
- algorithmic differentiation, gradients, Hessians, Taylor series, slopes
- finding all roots of a nonlinear system
- Galois field toolbox
- etc.

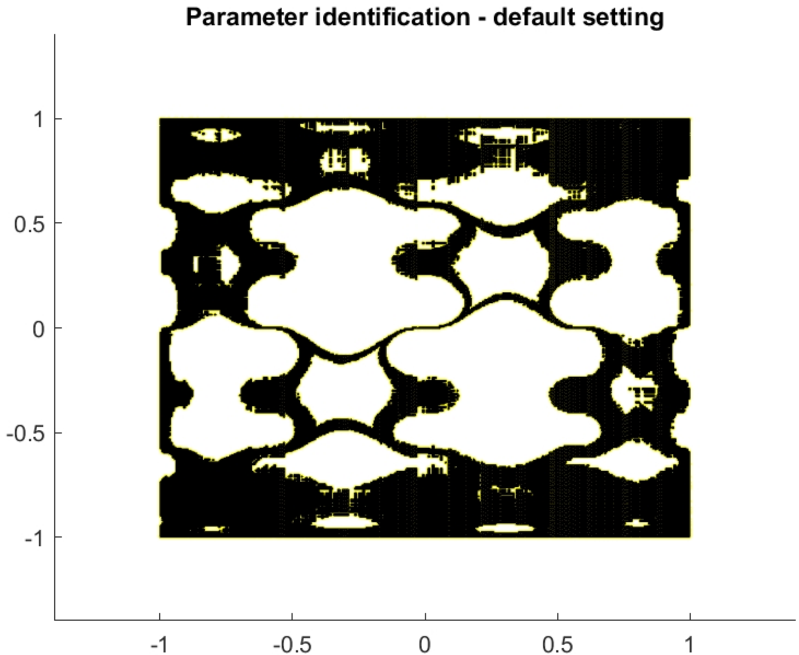
`https://www.tuhh.de/ti3/rump/intlab/`

Back
Close

Parameter identification - ordinary interval arithmetic

$$f = -(5y - 20y^2 + 16y^5)^6 + (-(5x - 20x^3 + 16x^5)^3 + 5y^2 - 20y^3 + 16y^5)^2$$

```
X = infsup(-1,1)*ones(2,1);
verifynlssparam(f,0,X)
verifynlssparamset('Display','~'));
```



Interval arithmetic is no panacea — The wrapping effect

Ordinary interval arithmetic (executable INTLAB code):

```
A = infsup(1,3); intDiff = A-A

intval intDiff =
[   -2.0000,    2.0000]
```

Interval arithmetic is no panacea — The wrapping effect

Ordinary interval arithmetic (executable INTLAB code):

```
A = infsup(1,3); intDiff = A-A

intval intDiff =
[   -2.0000,    2.0000]
```

Affine arithmetic (executable INTLAB code):

```
B = affari(infsup(1,3)); affDiff = B-B

affari affDiff =
[    0.0000,    0.0000]
```

Affine arithmetic - References

Ref.: Andrack, Comba, Stolfi 1994
Figueiredo/Stolfi, monograph 1997
Kashiwagi, monograph 2005
Stolfi, reference implementation 2007
R./Kashiwagi, Improvements of affine arithmetic, IEICE, 2015

Affine arithmetic - References

Ref.: Andrack, Comba, Stolfi 1994
Figueiredo/Stolfi, monograph 1997
Kashiwagi, monograph 2005
Stolfi, reference implementation 2007
R./Kashiwagi, Improvements of affine arithmetic, IEICE, 2015

Representation of affine quantities:

$$C := \langle c\,;\gamma\rangle = \{c + \sum_{i=1}^{k} \gamma_i \varepsilon_i : \ \varepsilon \in \mathcal{E}^k\} \ \text{ with } \mathcal{E} := [-1,1]$$

All ε_i vary independently in $\mathcal{E}$.

Affine arithmetic - References

Ref.: Andrack, Comba, Stolfi 1994
Figueiredo/Stolfi, monograph 1997
Kashiwagi, monograph 2005
Stolfi, reference implementation 2007
R./Kashiwagi, Improvements of affine arithmetic, IEICE, 2015

Representation of affine quantities:

$$C := \langle c\,;\gamma\rangle = \{c + \sum_{i=1}^{k} \gamma_i \varepsilon_i : \ \varepsilon \in \mathcal{E}^k\} \ \text{ with } \mathcal{E} := [-1,1]$$

All ε_i vary independently in $\mathcal{E}$.

For example,

```
A = affari(infsup(1,3));  B = affari(infsup(-2,4));
```

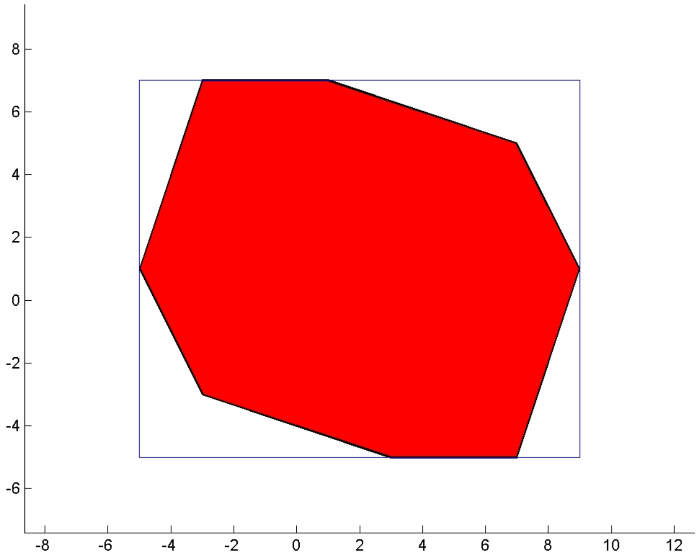
implies $A := \langle 2; 1\rangle$ and $B := \langle 1; 0, 3\rangle$

Back
Close

Affine arithmetic

Example: $C := \langle 2; 1, -2, 3, -1 \rangle$
$D := \langle 1; 3, 0, -1, 2 \rangle$

$$C \times D = \left\{ \begin{pmatrix} 2 + \varepsilon_1 - 2\varepsilon_2 + 3\varepsilon_3 - \varepsilon_4 \\ 1 + 3\varepsilon_1 - \varepsilon_3 + 2\varepsilon_4 \end{pmatrix} : \ \varepsilon_i \in [-1, 1] \right\}$$



Affine arithmetic improvements

Given $C := \langle c\,; \gamma \rangle \quad \hat{=} \quad c + \sum \gamma_i \varepsilon_i \quad \Rightarrow \quad f(C)$?

<u>Def.</u> $\quad f$ is represented by $[\![p, q, \Delta]\!]$ on $[a, b]$

$$\text{s.t. } \forall x \in \mathbf{X} : \quad |px + q - f(x)| \le \Delta$$

$\Rightarrow \quad$ Determine p, q, Δ for given f and given $[a, b]$

$\Rightarrow \quad$ Use $x \in [a, b] \quad \Leftrightarrow \quad x = c + \sum \gamma_i \varepsilon_i$ for $|\varepsilon_i| \le 1$

$$\Rightarrow \quad |pc + q + \sum p\gamma_i \varepsilon_i - f(x)| \le \Delta$$

i.e. $\quad \langle pc + q; p\gamma, \Delta \rangle$ represents $f(C)$ on $[a, b]$

Functions in the affine toolbox

sqrt, sqr,

exp, log, log2, log10, power,

sin, cos, tan, cot, sec, csc,

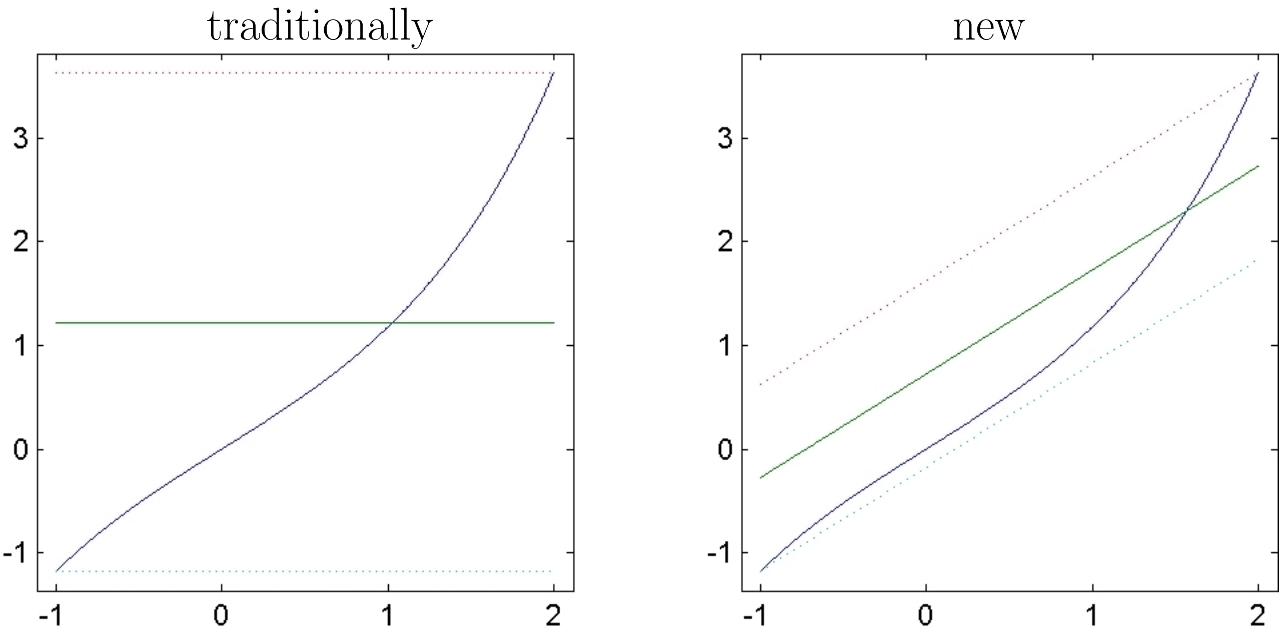
asin, acos, atan, acot, asec, acsc,

sinh, cosh, tanh, coth,

asinh, acosh, atanh, acoth,

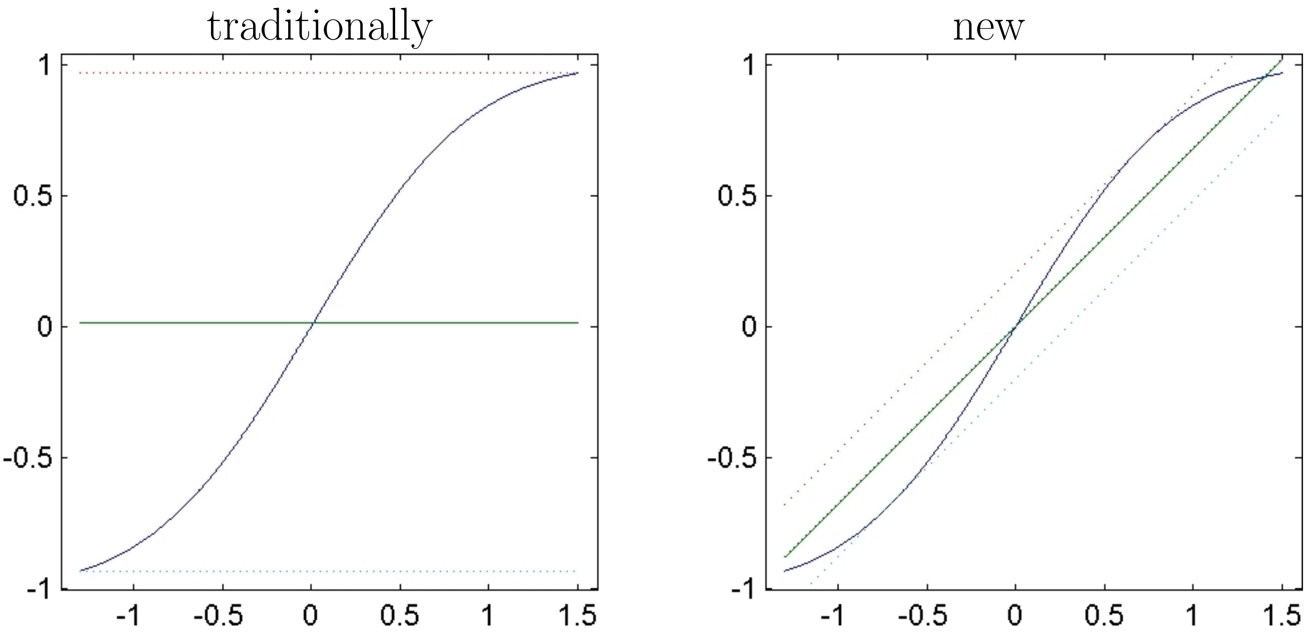
erf, erfc.

Chebyshev representation



Min-Range representation of sinh(x) on [-1,2]

Chebyshev representation

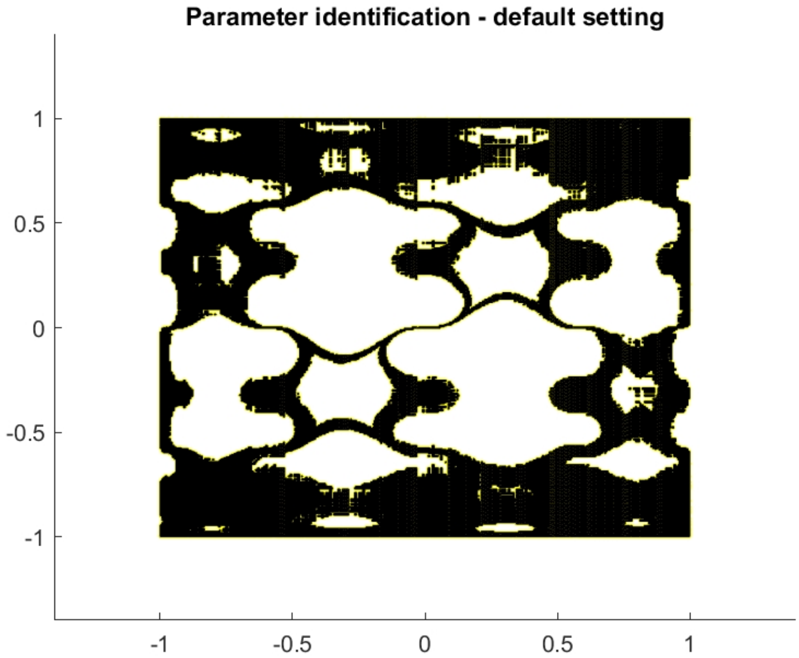


Chebyshev representation of erf(x) on [-1.3,1.5]

Parameter identification - ordinary interval arithmetic

$$f = -(5y - 20y^2 + 16y^5)^6 + (-(5x - 20x^3 + 16x^5)^3 + 5y^2 - 20y^3 + 16y^5)^2$$

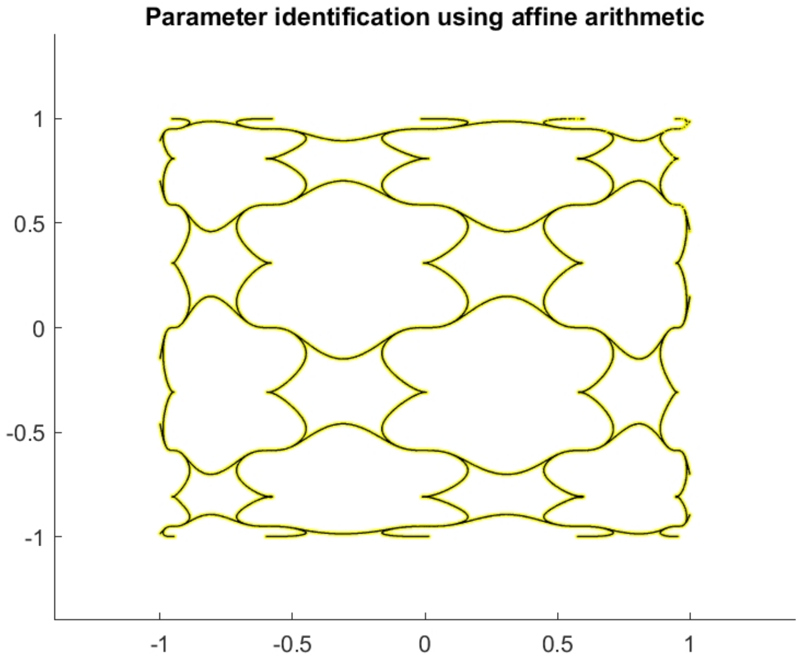
```
X = infsup(-1,1)*ones(2,1);
verifynlssparam(f,0,X)
verifynlssparamset('Display','~'));
```



Parameter identification - using affine arithmetic

$$f = -(5y-20y^2+16y^5)^6+(-(5x-20x^3+16x^5)^3+5y^2-20y^3+16y^5)^2$$

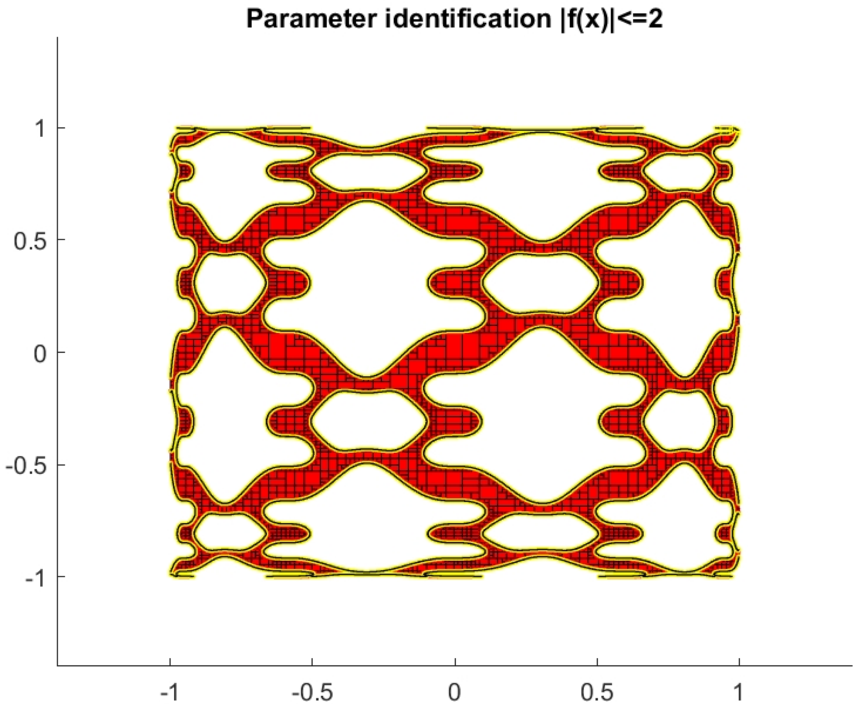
```
X = infsup(-1,1)*ones(2,1);
verifynlssparam(f,0,X)
verifynlssparamset('Display','~','Method','affari'));
```



Parameter identification - affine arithmetic, nontrivial interior

$$f = -(5y - 20y^2 + 16y^5)^6 + (-(5x - 20x^3 + 16x^5)^3 + 5y^2 - 20y^3 + 16y^5)^2$$

```
verifynlssparam(f,infsup(-0.2,0.2),X, ...
verifynlssparamset('Display','~','Method','affari'));
```



Limits of verification methods and of arithmetic

Verification methods can only solve well-posed problems

- No real interval can be verified to contain a root of $x^2 + 2x + 1$

- A complex interval can be verified to contain 2 roots

Limits of verification methods and of arithmetic

Verification methods can only solve well-posed problems

- No real interval can be verified to contain a root of $x^2 + 2x + 1$
- A complex interval can be verified to contain 2 roots

- We cannot verify that a matrix is singular
- We can verify that a matrix is nonsingular [even for $\mathrm{cond}(A) > 10^{100}$]

Limits of verification methods and of arithmetic

Verification methods can only solve well-posed problems

- No real interval can be verified to contain a root of $x^2 + 2x + 1$
- A complex interval can be verified to contain 2 roots

- We cannot verify that a matrix is singular
- We can verify that a matrix is nonsingular [even for $\mathrm{cond}(A) > 10^{100}$]

- An eigenvector to a double eigenvalue cannot be included

Limits of verification methods and of arithmetic

Verification methods can only solve well-posed problems

- No real interval can be verified to contain a root of $x^2 + 2x + 1$
- A complex interval can be verified to contain 2 roots

- We cannot verify that a matrix is singular
- We can verify that a matrix is nonsingular [even for $\mathrm{cond}(A) > 10^{100}$]

- An eigenvector to a double eigenvalue cannot be included

- $\sin\,(4\ \mathrm{atan}(1)) \geq 0$ cannot be decided in any precision

The van der Pol equation

$$y'' - c(1 - y^2)y' + y = 0 \quad \text{for some scalar c} > 0$$

rewritten into a system of first order ODEs

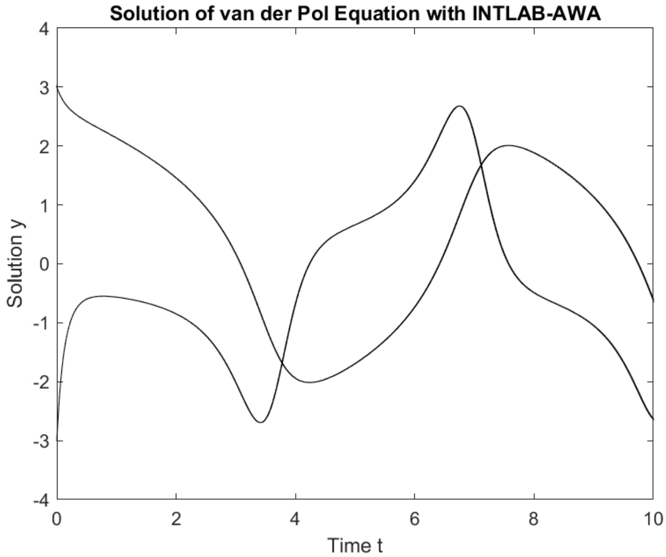
$$\begin{aligned} y_1' &= y_2 \ , \\ y_2' &= c(1 - y_1^2)y_2 - y_1 \end{aligned}$$

Initial conditions $y_0 \in \begin{pmatrix} 3 \\ -3 \end{pmatrix} \pm 0.001$

Solution of the van der Pol equation for $t \in [0, 10]$

```
function J = vdp_jac(t,y)
J = typeadjust([0,1;0,0],y);
J(2,1) = -2.*y(1).*y(2) - 1;
J(2,2) = 1 - sqr(y(1));

[T,Y] = awa(@vdp_fun,@vdp_jac,[0,10],midrad([3;-3],1e-3));
plot(T,Y)
```



Taylor models, implemented by Florian Bünger

K. Makino, Rigorous analysis of nonlinear motion in particle accelerators, Dissertation at Michigan State University, 1998

K. Makino and M. Berz, Suppression of the wrapping effect by Taylor model - based validated integrators, MSU HEP Report 40910, 2003

A. Neumaier, Taylor Forms – Use and Limits, Reliable Computing 9, pp. 43-79, 2003

F. Bünger, Shrink wrapping for Taylor models revisited, Numerical Algorithms 78(4), pp. 1001-1017, 2018

F. Bünger, Reducing the truncation error in Taylor model multiplication, accepted for publication, 2023

Definition of Taylor models

$$p(x) = \sum_{a,|a|\le d} p_a x^a, \quad |a| := a_1 + \ldots + a_n, \quad x^a := x_1^{a_1} \cdots x_n^{a_n},$$

on $D = [u_1, v_1] \times \cdots \times [u_n, v_n]$

An inclusion of $p(D-c)+E = \{p(x-c)+e \mid x \in D, e \in E\}$ is computed

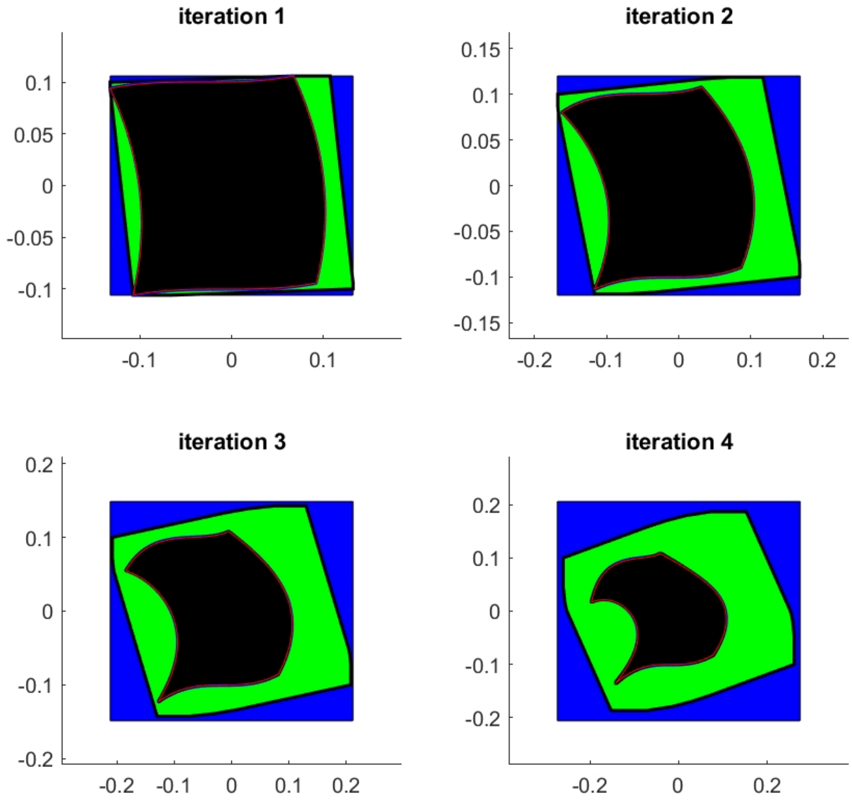
Mostly the standard domain $Ds := [-1,1]^n$, $cs := (0,\ldots,0)$ is used

In contrast to affine arithmetic, Taylor models need not to be convex

Options QR preconditiong, shrink wrapping,

Taylor models - An example

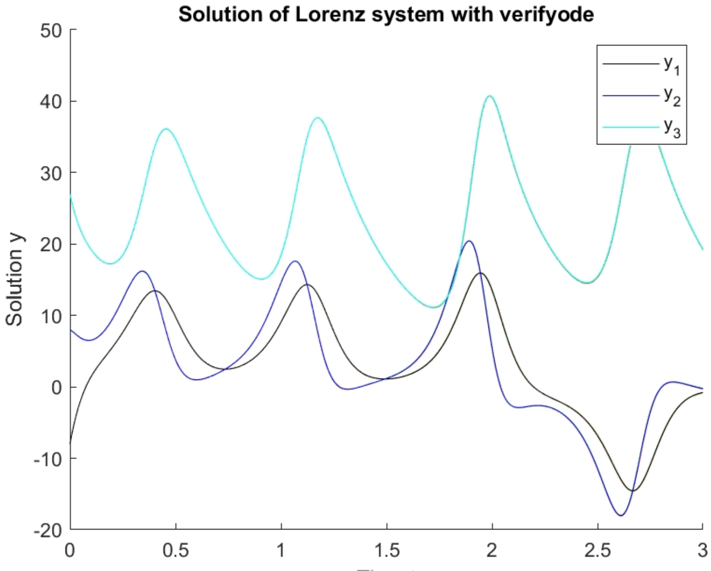
For $f(x, y) = (x - y(0.125 + 2y), y + 6x^3)$ compute the iterated image of $B := [-0.1, 0.1]^2$, i.e. $f(f(\ldots f(B)\ldots))$



Taylor models - The Lorenz system

$$\begin{aligned} y_1' &= \sigma(y_2 - y_1) \\ y_2' &= (\rho - y(3))y(1) - y(2) \qquad \text{for } \ \sigma = 10,\ \rho = 28,\ \beta = 8/3 \\ y_3' &= y_1 y_2 - \beta y_3 \end{aligned}$$

$$y_0 \in [-8.001, -7.998] \times [7.998, 8.001] \times [26.998, 27.001]$$



Summary

- Error standard models in numerical analysis
- Optimal bounds for IEEE-754 $+, -, \times, /, \sqrt{\cdot}$
- Weak sufficient assumptions for linearized bounds
- Optimal bounds for the error of summation
- error-free transformations
- Provably mathematical correct results
- Global optimization
- Affine arithmetic
- Taylor models

On verification methods:

S.M. Rump. Verification methods: Rigorous results using floating-point arithmetic. Acta Numerica, 19:287–449, 2010.

An application of affine arithmetic — Julia sets

$z_0, c \in \mathbb{C}: \qquad z_{k+1} := z_k^2 + c \quad \text{for } k \geq 1$

Given $c \in \mathbb{C}$, for which $z_0 \in \mathbb{C}$ is ∞ point of attraction?

Divergent for $\min\{|\mathrm{Re}\ c|, |\mathrm{Im}\ c|\} \geq 2$.

Color code	red	∞ point of attraction
	black	iteration bounded for all $k \geq 1$
	yellow	don't know