Can you count on your computer?

inspired from Nick Higham, Jean-Michel Muller, Siegfried Rump, Arnaud Tisserand and many others

NSV-3, 15 July 2010

Computing without errors is difficult...

I do hate sums. There is no greater mistake than to call arithmetic an exact science.
There are permutations and aberrations discernible to minds entirely noble like mine; subtle variations which ordinary accountants fail to discover; hidden laws of number which it requires a mind like mine to perceive.
For instance, if you add a sum from the bottom up, and then from the top down, the result is always different.

Maria La Touche, 1850

quoted in the Mathematical Gazette, volume 12, 1924, originally published in *The letters of a noble woman: Mrs. La Touche of Harristown*, 1908.

Summing from the bottom up, and then from the top down

Try summing, in radix 10 and with 3 digits for the mantissa, the numbers: (inspired from Ogita, Rump and Oishi (SISC 2005))

$$
\begin{array}{rr|l}
1\,0\,2 & & \times 10^{3} \cdot \\
5\,0\,0 & & \times 10^{0} \\
-\,1\,0\,2 & & \times 10^{3} \\
5\,0\,0 & & \times 10^{0} \\
 & .6\,7\,8 & \times 10^{-3} \\
-\quad 1\,0\,0 & & \times 10^{1}
\end{array}
$$

(rounding to nearest even)

Exact result: 0.678

Summing from the bottom up, and then from the top down

Try summing, in radix 10 and with 3 digits for the mantissa, the numbers: (inspired from Ogita, Rump and Oishi (SISC 2005))

$$
\begin{array}{rr|l}
1\,0\,2 & & \times 10^{3} \cdot \\
5\,0\,0 & & \times 10^{0} \\
-\,1\,0\,2 & & \times 10^{3} \\
5\,0\,0 & & \times 10^{0} \\
 & .6\,7\,8 & \times 10^{-3} \\
-\quad 1\,0\,0 & & \times 10^{1}
\end{array}
$$

(rounding to nearest even)

Exact result: 0.678

Summing from the bottom up

Try summing, in radix 10 and with 3 digits for the mantissa, the numbers:

1 0 2			$\times 10^{3}$			0	$\times 10^{0}$
	5 0 0		$\times 10^{0}$	-	1 0 2		$\times 10^{3}$
-	1 0 2		$\times 10^{3}$	-	1 0 2		$\times 10^{3}$
	5 0 0		$\times 10^{0}$	-	5 0 0		$\times 10^{0}$
		.6 7 8	$\times 10^{-3}$	-	1 0 0		$\times 10^{1}$
-	1 0 0		$\times 10^{1}$	-	1 0 0		$\times 10^{1}$

(rounding to nearest even)

Exact result: 0.678

Summing then from the top down

Try summing, in radix 10 and with 3 digits for the mantissa, the numbers:

	1 0 2		$\times 10^{3}$.		1 0 2		$\times 10^{3}$.
	5 0 0		$\times 10^{0}$		1 0 2		$\times 10^{3}$
-	1 0 2		$\times 10^{3}$		0		$\times 10^{0}$
	5 0 0		$\times 10^{0}$		5 0 0		$\times 10^{0}$
		.6 7 8	$\times 10^{-3}$		5 0 1		$\times 10^{0}$
-	1 0 0		$\times 10^{1}$	-	4 9 9		$\times 10^{0}$

(rounding to nearest even)

Exact result: 0.678 and **result computed from the bottom up: 0**.

Program "sum.c"

Provides a similar example, in base 2^{53} which is the width of the mantissa in double-precision floating-point arithmetic.
Summands cover 3 "windows":

- summands 0 and 2 cancel on the leftmost window,
- summands 1, 3 and 5 cancel in the intermediate window,
- summand 4 corresponds to the exact result, in the rightmost window.

Example built using ideas and algorithm 6.1 of *Accurate sum and dot product*, by T. Ogita, S.M. Rump, and S. Oishi. SIAM Journal on Scientific Computing (SISC), 26(6):19551988, 2005.

Can You "Count" on Your Computer? up to 6... (N. Higham)

$$2 - 1$$

$$\left(\frac{1}{\cos(100\pi+\pi/4)}\right)^2$$

$$3.0 * (tan(atan(10000000.0))/10000000.0)$$

$$\left(\cdots(\sqrt{\cdots\sqrt{4}})^2\cdots\right)^2 \text{ (20 fois)}$$

$$5 \times \frac{(1+e^{-100})-1}{(1+e^{-100})-1}$$

$$\frac{\ln(e^{6000})}{1000}$$

One, two, three, . . .

$2 - 1$	1.0000000000000000
$\left(\frac{1}{\cos(100\pi+\pi/4)}\right)^2$	2.0000000000001110
$3.0 * \frac{tan(atan(10000000.0))}{10000000.0}$	2.9999999998627116

Is this a FLOP?

. . . , four, five, six

$\left(\cdots(\sqrt{\cdots\sqrt{4}})^2\cdots\right)^2$ (20 fois)	4.0000000006294343
$5 \times \frac{(1+e^{-100})-1}{(1+e^{-100})-1}$	NaN
$\frac{\ln(e^{6000})}{1000}$	$+\infty$

One, two, three, . . .

- **One** $= 2 - 1$
  is exact: 1 and 2 are exactly representable and belong to the same "binad", thus the subtraction is exact.
- **Two** $= \left(\frac{1}{\cos(100\pi+\pi/4)}\right)^2$
  cannot be exact: π is not exactly representable, thus $100\pi + \pi/4$ is not exactly equal to $\pi/4 \mod 2\pi$, thus $\cos(100\pi + \pi/4)$ is not exactly $1/\sqrt{2}$ (which is anyway not exactly representable), thus its square is not exactly 2. Furthermore, quite often cos is not well specified (except in CRlibm).
- **Three** $= 3.0 * (\tan(\arctan(10000000.0))/10000000.0)$
  arctan of a large argument is very close to $\pi/2$, the floating-point evaluation of tan of an argument very close to $\pi/2$ can be very far from the exact value, because tan varies a lot, ie. a small error is amplified. In other words, tan(arctan) can be very far from the identity.

. . . , four, five, six (N. Higham)

- **Four** $= \left(\cdots (\sqrt{\cdots \sqrt{4}})^2 \cdots \right)^2$ **(20 fois)**
  Since the square root of 2 is not exactly representable, there is an error at each application of $\sqrt{}$; the result is very close to 1 (as for any square root iterated). The iterated squaring amplifies this error.
- **Five** $= 5 \times \frac{(1+e^{-100})-1}{(1+e^{-100})-1}$
  $\exp(-100)$ is very close to 0, $1 + \exp(-100)$ is rounded to 1, thus the denominator of the fraction is 0 and the result of a division by 0 is a NaN (Not a Number).
- **Six** $= \frac{\ln(e^{6000})}{1000}$
  When x is large, $\exp x$ overflows and $\ln(+\infty) = +\infty$.

Siegfried Rump's formula

$$
\begin{aligned}
a &= 77617 \\
b &= 33096 \\
&333.75b^6 + a^2 \cdot (11a^2 \cdot b^2 - b^6 + 121b^4 - 2) + 5.5b^8 + \frac{a}{2b}
\end{aligned}
$$

Two levels of cancellation:

- between terms of order of $a^2 \cdot b^6$ and b^8
- between terms of order of b^6, $a^4 \cdot b^2$, $a^2 \cdot b^4$

Exact result $\simeq 0.827396059946821368141165095479816$ ie. of order of $a/(2b)$ and the constant 2.

Siegfried Rump's formula

$$
\begin{aligned}
a &= 77617 \\
b &= 33096 \\
&333.75b^6 + a^2 \cdot (11a^2 \cdot b^2 - b^6 + 121b^4 - 2) + 5.5b^8 + \frac{a}{2b}
\end{aligned}
$$

Two levels of cancellation:

- between terms of order of $a^2 \cdot b^6$ and b^8
- between terms of order of b^6, $a^4 \cdot b^2$, $a^2 \cdot b^4$

Exact result $\simeq 0.827396059946821368141165095479816$ ie. of order of $a/(2b)$ and the constant 2.

Jean-Michel Muller's sequence

$$
\begin{aligned}
x_0 &= 4 \\
x_1 &= 17/4 \\
x_{n+1} &= 108 - \frac{815 - \frac{1500}{x_n}}{x_{n-1}}
\end{aligned}
$$

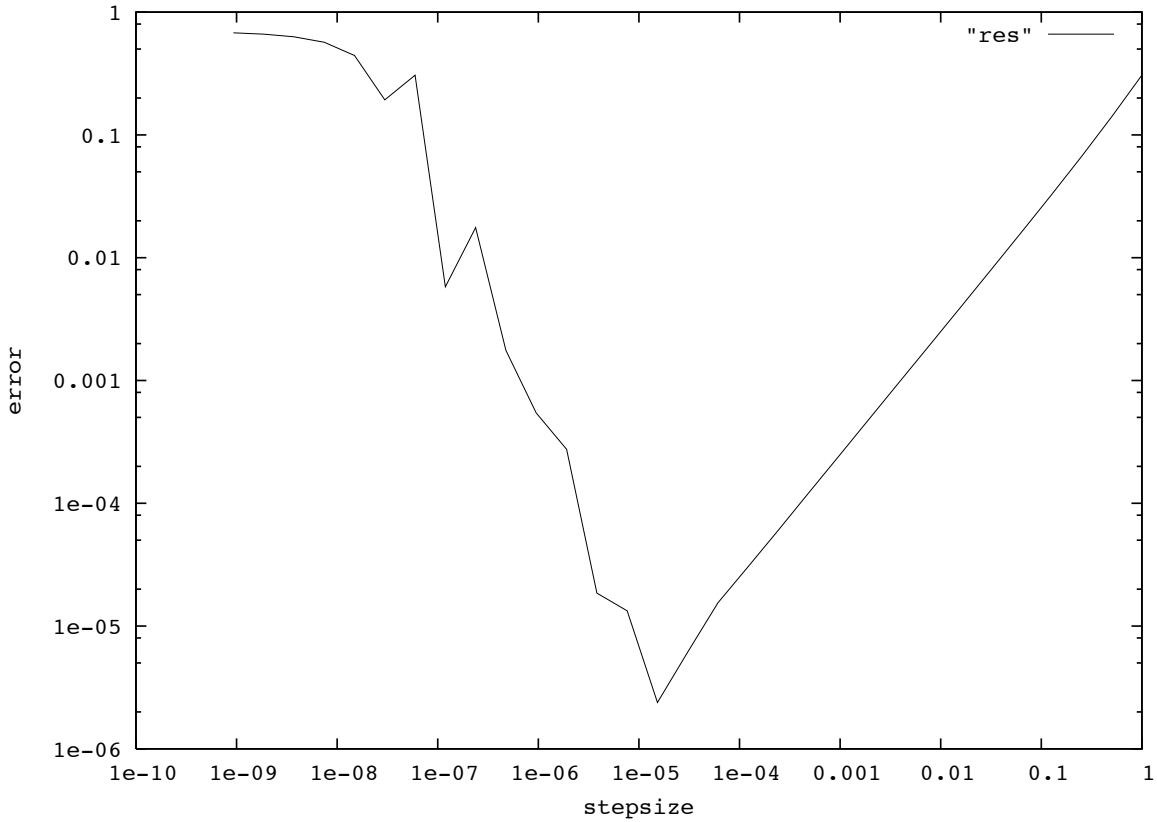
In theory: three fixed points (3, 5, 100),
3 and 5 are repellent, 100 is attractive.

With these initial values, the sequence should be on a trajectory converging to 5.
With roundoff errors, the computed trajectory converges to 100.

Numerical integration: $\int_1^2 \frac{dx}{x} = \ln 2$

using rectangles (based on the left point): $x_{n+1} = x_n + h/x_n$, for various choices of h the discretization step.
The method error is of order $\mathcal{O}(h^2)$.
When h is too large, the error is large also, the method error predominates.
When h is too small, the method error is small. . . but the accumulation of roundoff errors predominates.

Numerical integration: $\int_1^2 \frac{dx}{x} = \ln 2$



Absolute error in function of the stepsize.

What do I expect from the verification tools applied to these programs?

that they diagnose the loss of accuracy (due to the use of floating-point arithmetic), when there is one.