# Error continued.....

Department of Mathematics
IIT Guwahati

# Rude behavior of roundoff error!!

**Cancellation error:** how to avoid it in practice?

**Example** Suppose we wish to compute $y = \sqrt{x+1} - \sqrt{x}$ for $x = 100{,}000$ in a five-digit decimal arithmetic. Clearly, the number 100,001 cannot be represented in this floating point system exactly, and its representation in the system (when either chopping or rounding is used) is 100,000. In other words, for this value of $x$ in this floating point system, we have $x + 1 = x$. Thus, naively computing $\sqrt{x+1} - \sqrt{x}$ results in the value 0.

We can do much better if we use the identity

$$\frac{(\sqrt{x+1} - \sqrt{x})(\sqrt{x+1} + \sqrt{x})}{(\sqrt{x+1} + \sqrt{x})} = \frac{1}{\sqrt{x+1} + \sqrt{x}}.$$

Using this formula, computing the expression in 5 digit decimal arithmetic yields $1.5811 \times 10^{-3}.(Excersise)$ ( Exact value is $1.58113487 \times 10^{-3}$)

# Rude behavior of roundoff error!!

**Underflow and Overflow:** Consider a floating point system with $4$ decimal digits and $2$ exponent digits ( i.e. $-99 \leq e \leq 99,$). Compute

$$c = \sqrt{(a^2 + b^2)} \text{ for } a = 10^{60} \text{ and } b = 1.$$

$$fl(a) = 1.000 \times 10^{60}, \quad fl(b) = 1.000 \times 10^{0}.$$

$fl(a) \times fl(a) = 1.000 \times 10^{120}$ which cannot be represented in this floating system as exponent is three digit number.

# Rude behavior of roundoff error!!

**Underflow and Overflow:** Consider a floating point system with $4$ decimal digits and $2$ exponent digits ( i.e. $-99 \leq e \leq 99,$). Compute

$$c = \sqrt{(a^2 + b^2)} \text{ for } a = 10^{60} \text{ and } b = 1.$$

$$fl(a) = 1.000 \times 10^{60}, \quad fl(b) = 1.000 \times 10^{0}.$$

$fl(a) \times fl(a) = 1.000 \times 10^{120}$ which cannot be represented in this floating system as exponent is three digit number.
**How to avoid it?** Rewrite the expression as

$$c = s\sqrt{(a/s)^2 + (b/s)^2} \text{for any } s \neq 0.$$

Thus if we use $s = a = 10^{60}$ there will be an underflow in $(b/s)^2$, $fl((b/s)^2) = 1.000 \times 10^{-120}$, this will be set as zero. Finally we get the correct answer up to the precision of floating point system.

# Rude behavior of roundoff error!!

**Accumulation of roundoff error:** $x, y$ and $z$ are machine numbers in $32$ bit system.

$$
\begin{aligned}
fl[x(y+z)] &= [x\,fl(y+z)](1+\delta_1) \qquad |\delta_1| \leqq 2^{-24} \\
&= [x(y+z)(1+\delta_2)](1+\delta_1) \qquad |\delta_2| \leqq 2^{-24} \\
&= x(y+z)(1+\delta_2+\delta_1+\delta_2\delta_1) \\
&\approx x(y+z)(1+\delta_1+\delta_2) \\
&= x(y+z)(1+\delta_3) \qquad |\delta_3| \leqq 2^{-23}
\end{aligned}
$$

# Rude behavior of roundoff error!!

**Accumulation of roundoff error:** $x, y$ and $z$ are machine numbers in $32$ bit system.

$$
\begin{aligned}
fl[x(y+z)] &= [x\,fl(y+z)](1+\delta_1) \qquad |\delta_1| \leq 2^{-24} \\
&= [x(y+z)(1+\delta_2)](1+\delta_1) \qquad |\delta_2| \leq 2^{-24} \\
&= x(y+z)(1+\delta_2+\delta_1+\delta_2\delta_1) \\
&\approx x(y+z)(1+\delta_1+\delta_2) \\
&= x(y+z)(1+\delta_3) \qquad |\delta_3| \leq 2^{-23}
\end{aligned}
$$

## Theorem

*(On the relative roundoff error in adding) Let $x_0, x_1, \cdots, x_n$ be positive machine numbers in a computer whose unit roundoff error is $\eta$. Then the relative roundoff error in computing*

$$\sum_{i=0}^{n} x_i$$

*in the usual way is at most $(1+\eta)^n - 1 \approx n\eta$.*

# Rude behavior of roundoff error!!

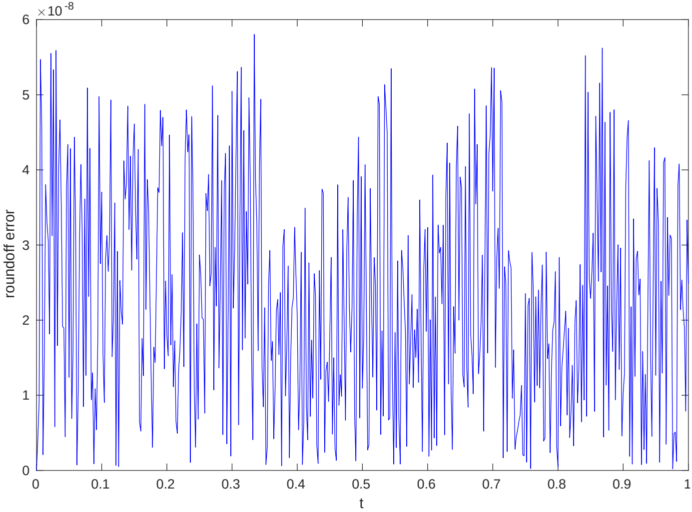


Figure: Error in sampling $exp(-t)(\sin(2\pi t) + 2)$ in single precision.

# Rude behavior of roundoff error!!

**Root of quadratic polynomial:** Example. Fining roots of $x^2 + 62.10x + 1 = 0$ in $4$ digits decimal floating point system. Exact roots are found to be

$$x_1 = -0.01610723 \quad \text{and} \quad x_2 = -62.08390.$$

$$\begin{aligned}\sqrt{b^2 - 4ac} &= \sqrt{(62.10)^2 - (4.000)(1.000)(1.000)} \\ &= \sqrt{3856. - 4.000} = \sqrt{3852.} = 62.06,\end{aligned}$$

$$fl(x_1) = \frac{-62.10 + 62.06}{2.000} = \frac{-0.04000}{2.000} = -0.02000,$$

$$\frac{|-0.01610723 + 0.02000|}{|-0.01610723|} \approx 2.4 \times 10^{-1}.$$

$$fl(x_2) = \frac{-62.10 - 62.06}{2.000} = \frac{-124.2}{2.000} = -62.10$$

$$\frac{|-62.0839 + 62.10|}{|-62.0839|} \approx 3.2 \times 10^{-4}.$$

# Rude behavior of roundoff error!!

[Compare with $ax^2 + bx + c = 0$] The effect is due to the fact that

$b^2 \gg 4ac$ and $\sqrt{b^2 - 4ac} \approx |b|$. To reduce the error we use the formula $x_1 x_2 = 1$ We compute $x_2$ as it is and

$$x_1 = 1/x_2; \quad fl(x_1) = 1.000/(-62.0) = -0.01610$$

$$\frac{|-0.01610723 + 0.01610|}{|-0.01610723|} \approx 4.4 \times 10^{-4}.$$

# Rude behavior of roundoff error!!

## Nested Arithmetic

Accuracy loss due to round-off error can also be reduced by rearranging calculations, as shown in the next example.

Evaluate $f(x) = x^3 - 6.1x^2 + 3.2x + 1.5$ at $x = 4.71$ using three-digit arithmetic.

| | $x$ | $x^2$ | $x^3$ | $6.1x^2$ | $3.2x$ |
|---|---|---|---|---|---|
| Exact | 4.71 | 22.1841 | 104.487111 | 135.32301 | 15.072 |
| Three-digit (chopping) | 4.71 | 22.1 | 104. | 134. | 15.0 |
| Three-digit (rounding) | 4.71 | 22.2 | 105. | 135. | 15.1 |

# Rude behavior of roundoff error!!

Exact: $f(4.71) = 104.487111 - 135.32301 + 15.072 + 1.5 = -14.263899.$

Three-digit (chopping): $f(4.71) = ((104. - 134.) + 15.0) + 1.5 = -13.5,$

Three-digit (rounding): $f(4.71) = ((105. - 135.) + 15.1) + 1.5 = -13.4.$

# Rude behavior of roundoff error!!

$$\text{Chopping: } \left|\frac{-14.263899+13.5}{-14.263899}\right| \approx 0.05, \quad \text{and Rounding: } \left|\frac{-14.263899+13.4}{-14.263899}\right| \approx 0.06.$$

# Rude behavior of roundoff error!!

$$\text{Chopping: } \left| \frac{-14.263899 + 13.5}{-14.263899} \right| \approx 0.05, \quad \text{and Rounding: } \left| \frac{-14.263899 + 13.4}{-14.263899} \right| \approx 0.06.$$

**Nested arithmetic:**

$$p_m(x) = c_0 + c_1 x + \cdots + c_n x^n \to p_n(x) = (\ldots((c_n x_n + c_{n-1})x + c_{n-2})x \ldots\ldots)x + c_0.$$

# Rude behavior of roundoff error!!

Chopping: $\left|\dfrac{-14.263899+13.5}{-14.263899}\right| \approx 0.05$, and Rounding: $\left|\dfrac{-14.263899+13.4}{-14.263899}\right| \approx 0.06.$

## Nested arithmetic:

$$p_m(x) = c_0 + c_1 x + \cdots + c_n x^n \rightarrow p_n(x) = (\ldots((c_n x_n + c_{n-1})x + c_{n-2})x \ldots\ldots)x + c_0.$$

$$f(x) = x^3 - 6.1x^2 + 3.2x + 1.5 = ((x - 6.1)x + 3.2)x + 1.5.$$

Using three-digit chopping arithmetic now produces

$$\begin{aligned} f(4.71) &= ((4.71 - 6.1)4.71 + 3.2)4.71 + 1.5 = ((-1.39)(4.71) + 3.2)4.71 + 1.5 \\ &= (-6.54 + 3.2)4.71 + 1.5 = (-3.34)4.71 + 1.5 = -15.7 + 1.5 = -14.2. \end{aligned}$$


$$f(x) = x^3 - 6.1x^2 + 3.2x + 1.5 = ((x - 6.1)x + 3.2)x + 1.5.$$

Using three-digit chopping arithmetic now produces

$$\begin{aligned} f(4.71) &= ((4.71 - 6.1)4.71 + 3.2)4.71 + 1.5 = ((-1.39)(4.71) + 3.2)4.71 + 1.5 \\ &= (-6.54 + 3.2)4.71 + 1.5 = (-3.34)4.71 + 1.5 = -15.7 + 1.5 = -14.2. \end{aligned}$$

# Rude behavior of roundoff error!!

$$\text{Three-digit (chopping):} \quad \left| \frac{-14.263899 + 14.2}{-14.263899} \right| \approx 0.0045;$$

$$\text{Three-digit (rounding):} \quad \left| \frac{-14.263899 + 14.3}{-14.263899} \right| \approx 0.0025.$$

# Rude behavior of roundoff error!!

$$\text{Three-digit (chopping):} \quad \left| \frac{-14.263899 + 14.2}{-14.263899} \right| \approx 0.0045;$$

$$\text{Three-digit (rounding):} \quad \left| \frac{-14.263899 + 14.3}{-14.263899} \right| \approx 0.0025.$$

You save number of operations! $\rightarrow$ the number of operations comes down to $\mathcal{O}(n)$ from $\mathcal{O}(n^2)$ when we calculate the value of a $n^{th}$ order polynomial.;

# Catastrophe due to roundoff error!! :(

- Patriot missile failure in Dhahran, Saudi Arabia, on February 25, 1991, which resulted in 28 deaths.pause
- Intel Pentium flaw (1994):
  http://en.wikipedia.org/wiki/Pentium_FDIV_bug
  $A = 4195835.0, B = 3145727.0; A - (A/B) * B =?$ (five digit arithmetic)

# Catastrophe due to roundoff error!! :(

- Patriot missile failure in Dhahran, Saudi Arabia, on February 25, 1991, which resulted in 28 deaths.pause
- Intel Pentium flaw (1994):
  http://en.wikipedia.org/wiki/Pentium_FDIV_bug
  $A = 4195835.0, B = 3145727.0; A - (A/B) * B = ?$ (five digit arithmetic)
- The Explosion of the Ariane 5, June 4- 1996.

# Numerical instability

To compute $(\frac{1}{3})^{15}$.

# Numerical instability

To compute $(\frac{1}{3})^{15}$.

$$x_0 = 1, \quad x_1 = \frac{1}{3}, \cdots \quad x_n = \frac{13}{3}x_{n-1} - \frac{4}{3}x_{n-2}.$$

Thus $x_{15} = (\frac{1}{3})^{15}$ can be computed successively using the sequence above.

# Numerical instability

To compute $(\frac{1}{3})^{15}$.

$$x_0 = 1, \quad x_1 = \frac{1}{3}, \cdots \quad x_n = \frac{13}{3} x_{n-1} - \frac{4}{3} x_{n-2}.$$

Thus $x_{15} = (\frac{1}{3})^{15}$ can be computed successively using the sequence above.
If there is no roundoff error in the computer we will get the exact value of $x_{15}$.

# Numerical instability

To compute $(\frac{1}{3})^{15}$.

$$x_0 = 1, \quad x_1 = \frac{1}{3}, \cdots \quad x_n = \frac{13}{3}x_{n-1} - \frac{4}{3}x_{n-2}.$$

Thus $x_{15} = (\frac{1}{3})^{15}$ can be computed successively using the sequence above.
If there is no roundoff error in the computer we will get the exact value of $x_{15}$.
What happens if there is roundoff error?

# Numerical instability

$$
\begin{aligned}
x_0 &= 1.0000000 \\
x_1 &= 0.3333333 && \text{(7 correctly rounded significant digits)} \\
x_2 &= 0.1111112 && \text{(6 correctly rounded significant digits)} \\
x_3 &= 0.0370373 && \text{(5 correctly rounded significant digits)} \\
x_4 &= 0.0123466 && \text{(4 correctly rounded significant digits)} \\
x_5 &= 0.0041187 && \text{(3 correctly rounded significant digits)} \\
x_6 &= 0.0013857 && \text{(2 correctly rounded significant digits)} \\
x_7 &= 0.0005131 && \text{(1 correctly rounded significant digit)} \\
x_8 &= 0.0003757 && \text{(0 correctly rounded significant digits)} \\
x_9 &= 0.0009437 \\
x_{10} &= 0.0035887 \\
x_{11} &= 0.0142927 \\
x_{12} &= 0.0571502 \\
x_{13} &= 0.2285939 \\
x_{14} &= 0.9143735 \\
x_{15} &= 3.657493 && \text{(incorrect with relative error of } 10^8\text{)}
\end{aligned}
$$

# Numerical instability

$$
\begin{aligned}
x_0 &= 1.0000000 \\
x_1 &= 0.3333333 && \text{(7 correctly rounded significant digits)} \\
x_2 &= 0.1111112 && \text{(6 correctly rounded significant digits)} \\
x_3 &= 0.0370373 && \text{(5 correctly rounded significant digits)} \\
x_4 &= 0.0123466 && \text{(4 correctly rounded significant digits)} \\
x_5 &= 0.0041187 && \text{(3 correctly rounded significant digits)} \\
x_6 &= 0.0013857 && \text{(2 correctly rounded significant digits)} \\
x_7 &= 0.0005131 && \text{(1 correctly rounded significant digit)} \\
x_8 &= 0.0003757 && \text{(0 correctly rounded significant digits)} \\
x_9 &= 0.0009437 \\
x_{10} &= 0.0035887 \\
x_{11} &= 0.0142927 \\
x_{12} &= 0.0571502 \\
x_{13} &= 0.2285939 \\
x_{14} &= 0.9143735 \\
x_{15} &= 3.657493 && \text{(incorrect with relative error of } 10^8\text{)}
\end{aligned}
$$

Speaking informally, we say that a numerical process is unstable if small errors made at one stage of the process are magnified in subsequent stages and seriously degrade the accuracy of the overall calculation.