Lecture # 03

Finite Digit & Nested Arithmetic

Finite Digit Arithmetic:

Example 3 Suppose that $x = \frac{5}{7}$ and $y = \frac{1}{3}$. Use five-digit chopping for calculating $x + y$, $x - y$, $x \times y$, and $x \div y$.

$$x = \frac{5}{7} = 0.\overline{714285} \quad \text{and} \quad y = \frac{1}{3} = 0.\overline{3}$$

$$x \oplus y = fl(fl(x) + fl(y)) = fl\left(0.71428 \times 10^0 + 0.33333 \times 10^0\right)$$
$$= fl\left(1.04761 \times 10^0\right) = 0.10476 \times 10^1.$$

Error Analysis:

The true value is $x + y = \frac{5}{7} + \frac{1}{3} = \frac{22}{21}$, so we have

$$\text{Absolute Error} = \left| \frac{22}{21} - 0.10476 \times 10^{1} \right| = 0.190 \times 10^{-4}$$

and

$$\text{Relative Error} = \left| \frac{0.190 \times 10^{-4}}{22/21} \right| = 0.182 \times 10^{-4}.$$

Operation	Result	Actual value	Absolute error	Relative error
$x \oplus y$	0.10476×10^{1}	22/21	0.190×10^{-4}	0.182×10^{-4}
$x \ominus y$	0.38095×10^{0}	8/21	0.238×10^{-5}	0.625×10^{-5}
$x \otimes y$	0.23809×10^{0}	5/21	0.524×10^{-5}	0.220×10^{-4}
$x \oslash y$	0.21428×10^{1}	15/7	0.571×10^{-4}	0.267×10^{-4}

Loss of Significance:

Let $p = 0.54617$ and $q = 0.54601$. Use four-digit arithmetic to approximate $p - q$ and determine the absolute and relative errors using **(a)** rounding and **(b)** chopping.

Solution The exact value of $r = p - q$ is $r = 0.00016$.

(a) Suppose the subtraction is performed using four-digit rounding arithmetic. Rounding p and q to four digits gives $p^* = 0.5462$ and $q^* = 0.5460$, respectively, and $r^* = p^* - q^* = 0.0002$ is the four-digit approximation to r. Since

$$\frac{|r - r^*|}{|r|} = \frac{|0.00016 - 0.0002|}{|0.00016|} = 0.25,$$

the result has only one significant digit, whereas p^* and q^* were accurate to four and five significant digits, respectively.

(b) If chopping is used to obtain the four digits, the four-digit approximations to p, q, and r are $p^* = 0.5461$, $q^* = 0.5460$, and $r^* = p^* - q^* = 0.0001$. This gives

$$\frac{|r - r^*|}{|r|} = \frac{|0.00016 - 0.0001|}{|0.00016|} = 0.375,$$

which also results in only one significant digit of accuracy. ■

Loss of Significance:

occurs in numerical calculations when too many significant digits cancel

$$\begin{array}{r} 123.4567 \\ -\ 123.4566 \\ \hline 000.0001 \end{array}$$

Remedy:

Example

Calculate $\sqrt{9.01} - 3$ on a three-decimal-digit

$$\sqrt{9.01} - 3 = \frac{(\sqrt{9.01} - 3)(\sqrt{9.01} + 3)}{\sqrt{9.01} + 3}$$

$$= \frac{9.01 - 3^2}{\sqrt{9.01} + 3}$$

$$= \frac{0.01}{3.00 + 3} = \frac{.01}{6} = 0.00167 \approx 1.67 \times 10^{-3}.$$

Remedies:

1. Rationalizing
2. Using Series Expansion
3. Using Trigonometric identities
4. Reformulation

Remedies:

Using Trigonometric identities

As a simple example, consider the function

$$f(x) = \cos^2(x) - \sin^2(x)$$

There will be loss of significance at $x = \pi/4$.

The problem can be solved by the simple substitution

$$\cos^2(x) - \sin^2(x) = cos(2x)$$

The quadratic formula states that the roots of $ax^2 + bx + c = 0$, when $a \neq 0$, are

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Consider this formula applied to the equation $x^2 + 62.10x + 1 = 0$, whose roots are approximately

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

use four-digit rounding arithmetic in the calculations to determine the root.

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

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

a poor approximation to $x_1 = -0.01611$, with the large relative error

$$\frac{|-0.01611 + 0.02000|}{|-0.01611|} \approx 2.4 \times 10^{-1}.$$

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

has the small relative error

$$\frac{|-62.08 + 62.10|}{|-62.08|} \approx 3.2 \times 10^{-4}.$$

To overcome loss of significance of x_1

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}\left(\frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}}\right) = \frac{b^2 - (b^2 - 4ac)}{2a(-b - \sqrt{b^2 - 4ac})},$$

which simplifies to an alternate quadratic formula

$$x_1 = \frac{-2c}{b + \sqrt{b^2 - 4ac}}. \tag{1.2}$$

Using (1.2) gives

$$fl(x_1) = \frac{-2.000}{62.10 + 62.06} = \frac{-2.000}{124.2} = -0.01610,$$

which has the small relative error 6.2×10^{-4}.

Nested Arithmetic:

Accuracy loss due to round-off error can also be reduced by rearranging calculations.

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

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

Nested Scheme for the same problem:

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

Using three-digit chopping arithmetic now produces

$$f(4.71) = ((4.71 - 6.1)4.71 + 3.2)4.71 + 1.5 = ((-1.39)(4.71) + 3.2)4.71 + 1.5$$
$$= (-6.54 + 3.2)4.71 + 1.5 = (-3.34)4.71 + 1.5 = -15.7 + 1.5 = -14.2.$$

In a similar manner, we now obtain a three-digit rounding answer of -14.3. The new relative errors are

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

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

Nesting has reduced the relative error for the chopping approximation to less than 10% of that obtained initially. For the rounding approximation the improvement has been even more dramatic; the error in this case has been reduced by more than 95%. □

Previously:

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

Do
Q. 1,2,11 & 13 from Ex # 1.1
Q.1,4,5,6,7,8 & 13 from Ex # 1.2