Ch 01-3: Error Analysis

Dr. Feras Fraige

The number p^* is said to approximate p to t **significant digits** (or figures) if t is the largest nonnegative integer for which

$$\frac{|p - p^*|}{|p|} \leq 5 \times 10^{-t}.$$

Table 1.1 illustrates the continuous nature of significant digits by listing, for the various values of p, the least upper bound of $|p - p^*|$, denoted max $|p - p^*|$, when p^* agrees with p to four significant digits.

Table 1.1

p	0.1	0.5	100	1000	5000	9990	10000
max $\|p - p^*\|$	0.00005	0.00025	0.05	0.5	2.5	4.995	5.

Finite-Digit Arithmetic

- Symbols for machine operation $\oplus, \ominus, \otimes, \oplus\!\!\!\!\div$

Assume that the floating-point representations $fl(x)$ and $fl(y)$ are given for the real numbers x and y and that the symbols $\oplus$, $\ominus$, $\otimes$, $\oplus\!\!\!\!\div$ represent machine addition, subtraction, multiplication, and division operations, respectively. We will assume a finite-digit arithmetic given by

$$x \oplus y = fl(fl(x) + fl(y)), \quad x \otimes y = fl(fl(x) \times fl(y)),$$

$$x \ominus y = fl(fl(x) - fl(y)), \quad x \oplus\!\!\!\!\div\, y = fl(fl(x) \div fl(y)).$$

This arithmetic corresponds to performing exact arithmetic on the floating-point representations of x and y and then converting the exact result to its finite-digit floating-point representation.

Rounding arithmetic is easily implemented in Maple. For example, the command

$Digits := 5$

causes all arithmetic to be rounded to 5 digits. To ensure that Maple uses approximate rather than exact arithmetic we use the *evalf*. For example, if $x = \pi$ and $y = \sqrt{2}$ then

$evalf(x);\ evalf(y)$

produces 3.1416 and 1.4142, respectively. Then $fl(fl(x) + fl(y))$ is performed using 5-digit rounding arithmetic with

$evalf(evalf(x) + evalf(y))$

which gives 4.5558. Implementing finite-digit chopping arithmetic is more difficult and requires a sequence of steps or a procedure. Exercise 27 explores this problem.

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 $x \div y$.

Solution Note that

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

implies that the five-digit chopping values of x and y are

$$fl(x) = 0.71428 \times 10^0 \quad \text{and} \quad fl(y) = 0.33333 \times 10^0.$$

Thus

$$\begin{aligned} x \oplus y &= fl(fl(x) + fl(y)) = fl\left(0.71428 \times 10^0 + 0.33333 \times 10^0\right) \\ &= fl\left(1.04761 \times 10^0\right) = 0.10476 \times 10^1. \end{aligned}$$

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

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

Table 1.2 lists the values of this and the other calculations.

Table 1.2

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}

The maximum relative error for the operations in Example 3 is 0.267×10^{-4}, so the arithmetic produces satisfactory five-digit results. This is not the case in the following example.

Example 4 Suppose that in addition to $x = \frac{5}{7}$ and $y = \frac{1}{3}$ we have

$$u = 0.714251, \quad v = 98765.9, \quad \text{and} \quad w = 0.111111 \times 10^{-4},$$

so that

$$fl(u) = 0.71425 \times 10^{0}, \quad fl(v) = 0.98765 \times 10^{5}, \quad \text{and} \quad fl(w) = 0.11111 \times 10^{-4}.$$

Determine the five-digit chopping values of $x \ominus u$, $(x \ominus u) \oplus w$, $(x \ominus u) \otimes v$, and $u \oplus v$.

Solution These numbers were chosen to illustrate some problems that can arise with finite-digit arithmetic. Because x and u are nearly the same, their difference is small. The absolute error for $x \ominus u$ is

$$\begin{aligned} |(x-u)-(x \ominus u)| &= |(x-u)-(fl(fl(x)-fl(u)))| \\ &= \left|\left(\frac{5}{7}-0.714251\right)-\left(fl\left(0.71428 \times 10^{0}-0.71425 \times 10^{0}\right)\right)\right| \\ &= \left|0.347143 \times 10^{-4}-fl\left(0.00003 \times 10^{0}\right)\right| = 0.47143 \times 10^{-5}. \end{aligned}$$

This approximation has a small absolute error, but a large relative error

$$\left|\frac{0.47143 \times 10^{-5}}{0.347143 \times 10^{-4}}\right| \leq 0.136.$$

The subsequent division by the small number w or multiplication by the large number v magnifies the absolute error without modifying the relative error. The addition of the large and small numbers u and v produces large absolute error but not large relative error. These calculations are shown in Table 1.3. ■

Table 1.3

Operation	Result	Actual value	Absolute error	Relative error
$x \ominus u$	0.30000×10^{-4}	0.34714×10^{-4}	0.471×10^{-5}	0.136
$(x \ominus u) \oslash w$	0.27000×10^{1}	0.31242×10^{1}	0.424	0.136
$(x \ominus u) \otimes v$	0.29629×10^{1}	0.34285×10^{1}	0.465	0.136
$u \oplus v$	0.98765×10^{5}	0.98766×10^{5}	0.161×10^{1}	0.163×10^{-4}