1. Types of Error

Let p be the true value, and suppose we use p^* to approximate p. The **actual error** is $p - p^*$ (note the order). The **absolute error** is $|p - p^*|$ (always non-negative). The **relative error** is $\frac{|p-p^*|}{|p|}$ (also always nonngative), provided that $p \neq 0$.

Example. The relative **round-off error** in approximating a real number y satisfies

$$error = \left|\frac{y - fl\,(y)}{y}\right| = \left|1 - \frac{fl\,(y)}{y}\right|.$$

Suppose we perform the k-digit chopping for decimal representation for the number

$$y = 0.d_1d_2\ldots d_kd_{k+1}d_{k+2}\ldots \times 10^n,$$

then the relative error is

$$\begin{aligned}
\left|\frac{y - fl\,(y)}{y}\right| &= \left|\frac{0.d_1d_2\ldots d_kd_{k+1}d_{k+2}\ldots \times 10^n - 0.d_1d_2\ldots d_k \times 10^n}{0.d_1d_2\ldots d_kd_{k+1}d_{k+2}\ldots \times 10^n}\right| \\
&= \left|\frac{0.\underbrace{0\ldots0}_{k \text{ zeros}}d_{k+1}d_{k+2}\ldots \times 10^n}{0.d_1d_2\ldots d_kd_{k+1}d_{k+2}\ldots \times 10^n}\right| \\
&= \left|\frac{0.d_{k+1}d_{k+2}\ldots \times 10^{n-k}}{0.d_1d_2\ldots d_kd_{k+1}d_{k+2}\ldots \times 10^n}\right| \\
&= \left|\frac{0.d_{k+1}d_{k+2}\ldots}{0.d_1d_2\ldots d_kd_{k+1}d_{k+2}\ldots}\right| \times 10^{-k}.
\end{aligned}$$

Since $d_1 \neq 0$, the minimal value of the denominator is 0.1. The numerator is bounded above by 1. Therefore

$$\left|\frac{y - fl\,(y)}{y}\right| \leqslant \frac{1}{0.1} \times 10^{-k} = 10^{-k+1}.$$

2. Significant Digits

We have all been asked to give a numerical answer up to certain number of significant digits. What's the proper mathematical definition?

Suppose you have obtain p as your true answer. It is repeated decimal number that is hard to present. Your professor asks you to provide p^* by recording k significant digits of p.

The number p^* is said to approximate p to k **significant digits** if k is the largest nonnegative integer for which

$$\left|\frac{p - p^*}{p}\right| \leqslant 5 \times 10^{-k}$$

Example. Sanity check for significant digits. Suppose we are looking at $p = \pi = 3.1415926535....$ Suppose now a certain algorithm yields

$$p^* = 3.141592234$$

as an approximation to π. Does the above definition make sense? Let's compute the relative error.

$$\begin{aligned}\left|\frac{p-p^*}{p}\right| &= \left|\frac{3.1415926535... - 3.141592234}{3.1415926535...}\right| \\ &= \frac{4.19589793... \times 10^{-7}}{3.1415926535...} \\ &= 1.33559579 \times 10^{-7} \\ &\leqslant 5 \times 10^{-7}\end{aligned}$$

Now, is $k = 7$ the largest nonnegative integer that satisfies the inequality? Let's try $k = 8$. We note that certainly $1.33559579 \times 10^{-7} \nleqslant 5 \times 10^{-8}$ (actually bigger than) – and thus, we go with $k = 7$. So, are we matching π with 7 significant digits?

$$\begin{aligned}\pi &= \underline{3.1415926}535... \\ p^* &= \underline{3.1415922}34\end{aligned}$$

yes, indeed!

Remark. Why is the bound 5×10^{-k}? Why 5 but not 6 or 4?

3. Finite-Digit Arithmetic

$$\begin{aligned}x \oplus y &= fl\left(fl\left(x\right) + fl\left(y\right)\right) \\ x \ominus y &= fl\left(fl\left(x\right) - fl\left(y\right)\right) \\ x \otimes y &= fl\left(fl\left(x\right) \times fl\left(y\right)\right) \\ x \oslash y &= fl\left(fl\left(x\right) \div fl\left(y\right)\right)\end{aligned}$$

Example. Let $x = 2/3$ and $y = 4/7$. Compute the four elementary operations in finite-digit arithmetic.

We try addition first. The true answer is $s = x + y = \frac{26}{21}$ while $x = 0.\overline{6}$ and $y = 0.\overline{571428}$. Using finite-digit arithmetic, say, with 5-digit chopping, we have

$$\begin{aligned}s^* = x \oplus y &= fl\left(fl\left(x\right) + fl\left(y\right)\right) \\ &= fl\left(0.66666 \times 10^0 + 0.57142 \times 10^0\right) \\ &= fl\left(1.23808 \times 10^0\right) \\ &= fl\left(0.123808 \times 10^1\right) \\ &= 0.12380 \times 10^1.\end{aligned}$$

Now, how much did we miss?

$$Error_{abs} = |s - s^*| = \left|\frac{26}{21} - 0.12380 \times 10^1\right| = 9.52 \times 10^{-5},$$

and

$$Error_{rel} = \left|\frac{s - s^*}{s}\right| = \left|1 - \frac{0.12380 \times 10^1}{\frac{26}{21}}\right| = 7.69 \times 10^{-5}.$$

4. Catastrophic Cancellations

When two almost equal numbers are subtracted from one another, the floating-point representation of the answer may incur large relative error to the true answer, regardless of whether the individual representation of each number is at a high precision. In the following example, we prescribe a simple four-digit arithmetic to approximate the difference of two almost equal numbers.

Example. Let $p = 0.85655$ and $q = 0.85642$. Consider four-digit (a) chopping and (b) rounding. Find their absolute and relative errors.

The true answer is $r = p - q = 0.00013$.

(1) Chopping yields

$$\begin{aligned} r^* = p \ominus q &= fl\left(fl\left(p\right) - fl\left(q\right)\right) \\ &= fl\left(0.8565 - 0.8564\right) \\ &= fl\left(0.0001\right) \\ &= 0.1 \times 10^{-3}. \end{aligned}$$

Thus

$$Error_{abs} = |r - r^*| = \left|0.00013 - 0.1 \times 10^{-3}\right| = 3 \times 10^{-5}$$

which is not bad. However,

$$Error_{rel} = \left|\frac{r - r^*}{r}\right| = \frac{3 \times 10^{-5}}{0.00013} = \frac{3}{13} \approx 0.231 \leqslant 5 \times 10^{-1}$$

which means we are incurring a 23% relative error – something you can't overlook. The last step shows that the approximation and the true answer has one significant digit of accuracy.

(2) Rounding yields

$$\begin{aligned} r^* = p \ominus q &= fl\left(fl\left(p\right) - fl\left(q\right)\right) \\ &= fl\left(0.8566 - 0.8564\right) \\ &= fl\left(0.0002\right) \\ &= 0.2 \times 10^{-3}. \end{aligned}$$

Absolute error is still okay,

$$Error_{abs} = |r - r^*| = \left|0.00013 - 0.2 \times 10^{-3}\right| = 7 \times 10^{-5}.$$

But, relative error is even worse,

$$Error_{rel} = \left|\frac{r - r^*}{r}\right| = \frac{7 \times 10^{-5}}{0.00013} = \frac{7}{13} \approx 0.538 \leqslant 5 \times 10^{-0},$$

more than 50% relative error! The last step shows that the approximation vs true answer have zero significant digits in common! We are now really comparing apples to oranges.

In practice, the use of algebraic formulas to solve certain problems must be taken with care. A canonical example would be the quadratic formula.

$$x_+ = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad x_- = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Now, suppose $b > 0$ but $b^2 \gg 4ac$ (meaning that b^2 is far larger than $4ac$). An example of a quadratic equation would be

$$x^2 + 10^5 x + 1 = 0.$$

The negative root x_- is in fact okay, since we are in fact doing addition $-\left(b + \sqrt{b^2 - 4ac}\right)$ in the numerator. However, the positive is subject to catastrophic cancellation because

$$\sqrt{b^2 - 4ac} \approx \sqrt{b^2} = b$$

so $\sqrt{b^2 - 4ac}$ and b are pretty close to each other. As a result, we expect x_+ to be pretty small, and sometimes, so small that the rounding error eats up a chunk of it.

Example. Consider the quadratic equation $x^2 + 75x + 1 = 0$. True answers to a high precision are the following

$$x_+ = -0.01333570454, \quad x_- = -74.9866642955.$$

Let us use four-digit rounding in computing the roots using the quardratic formula.

$$\sqrt{b^2 - 4ac} = \sqrt{75^2 - 4} = \sqrt{5621} = 74.97.$$

Therefore,

$$fl\,(x_+) = \frac{-75 + 74.97}{2.00} = \frac{-0.03}{2} = -0.015.$$

This approximation of the root has relative error

$$\left| \frac{x_+ - (-0.015)}{x_+} \right| = 0.1248$$

about 12.5% relative error.

Solution 1. So, what would be the remedy to such atrocious relative error? We simply need to turn around the subtraction into addition. Blessed by the conjugate of the square root expression, we can rewrite

$$\begin{aligned} x_+ &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} \\ &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} \frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}} \\ &= -\frac{1}{2a} \frac{4ac}{b + \sqrt{b^2 - 4ac}} \\ &= \frac{-2c}{b + \sqrt{b^2 - 4ac}} \end{aligned}$$

Now, the denominator involves an innocent addition.

Solution 2. Another remedy would be utilizing the relationship between the two roots. In a quadratic equation, $x^2 + bx + c = 0$, we know

$$\begin{aligned} x_+ + x_- &= -b \\ x_+ x_- &= c \end{aligned}$$

Using the good root approximation x_-, we rewrite

$$x_+ = \frac{c}{x_-}$$

where we avoided any sort of subtraction.