Lecture `loss`: Loss of significance

September 2, 2022

Summary: First introducing the concept of significant digits, these notes describe why subtraction of nearly equal numbers is prone to loss of significance.

References: Part of what follows stems from the first edition of T. Sauer's textbook *Numerical Analysis*. It also relies on some PDF slides by C. Trenchea of the Department of Mathematics, University of Pittsburgh. Finally, some material is from `Wikipedia`.

Errors

Suppose the number x_A is an approximation to a "true" number $x_T \in \mathbb{R}$. (Do such true numbers really exist? Probably not, but it's a tremendously useful abstraction.) That is, x_A = approximate value, and x_T = true value. The *error* (or *absolute error*) of x_A relative to x_T is

$$\texttt{absolute error}(x_A) = x_T - x_A,$$

and the *relative error* is

$$\texttt{relative error}(x_A) = \frac{x_T - x_A}{x_T}.$$

These quantities can be negative, and often we use the same terms to refer to their magnitudes

$$\texttt{abserr}(x_A) = |x_T - x_A|, \qquad \texttt{relerr}(x_A) = \frac{|x_T - x_A|}{|x_T|}.$$

The relative error is not defined if $x_T = 0$, but it is typically a more meaningful measure. Indeed, the quantity $100 \times \texttt{relerr}(x_A)$ is the *percent error* of x_A relative to x_T. For example, suppose it's known that a metal bar weighs $x_T = 1.70\texttt{kg}$ (perhaps through the use of an accurate scale). A crude scale determines that $x_A = 1.45\texttt{kg}$. Then the percent error is

$$\frac{|1.70\texttt{kg} - 1.45\texttt{kg}|}{|1.70\texttt{kg}|} \times 100 \simeq 15\%.$$

Significant digits

Roughly, the number of *significant digits* (or *significant figures*) in x_A relative to x_T is the number of its leading digits that match those of x_T, although the last matching digit may not be counted in certain circumstances.

> *Precisely, if the error $|x_T - x_A|$ is ≤ 5 in the digit corresponding to the $(p+1)$st digit of x_T, counting left-to-right from the first nonzero digit of x_T, then we say that x_A has at least p significant digits of accuracy relative to x_T.*

Here is a diagram which illustrates the definition.

$$\begin{array}{rcl} x_T & = & (\overset{1}{d_N} \quad \cdots \quad \overset{N+1}{d_0} \bullet \overset{N+2}{d_{-1}} \quad \overset{N+3}{d_{-2}} \quad \cdots \quad \overset{p}{d_{-(p-N-1)}} \quad \overset{p+1}{d_{-(p-N)}} \quad \cdots) \\ |x_T - x_A| & = & (0 \quad \cdots \quad 0 \bullet 0 \quad 0 \quad \cdots \quad 0 \quad a_{-(p-N)} \quad \cdots) \end{array}$$

If the red digit $a_{-(p-N)} \leq 5$, then x_A has p significant digits relative to x_T; otherwise, we should color the 0 immediately to the left of $a_{-(p-N)}$ red, and x_A has $p-1$ significant digits relative to x_T.

For example, take $x_T = \pi = 3.14159265\cdots$ and $x_A = \frac{22}{7} = 3.1428571\cdots$. Then

$$\begin{array}{rcl} x_T & = & (3. \quad 1 \quad 4 \quad 1 \quad 5 \quad 9 \quad 2 \quad 6 \quad 5 \quad \cdots) \\ |x_T - x_A| & = & (0. \quad 0 \quad 0 \quad 1 \quad 2 \quad 6 \quad 4 \quad 4 \quad 8 \quad \cdots) \end{array}$$

and x_A has 3 significant digits. Now take $x_T = \frac{7}{9} = 0.7777777\cdots$ and $x_A = 0.7777$. Then

$$\begin{array}{rcl} x_T & = & (0. \quad 7 \quad 7 \quad 7 \quad 7 \quad 7 \quad 7 \quad 7 \quad 7 \quad \cdots) \\ |x_T - x_A| & = & (0. \quad 0 \quad 0 \quad 0 \quad 0 \quad 7 \quad 7 \quad 7 \quad 7 \quad\quad) \end{array}$$

Here, the first 7 appearing in $|x_T - x_A|$ is not marked in red since it's > 5. The number x_A has 3 significant digits. For a final example, take $x_T = 51.495$ and $x_A = 51.493$. Then

$$\begin{array}{rcl} x_T & = & (5 \quad 1. \quad 4 \quad 9 \quad 5) \\ |x_T - x_A| & = & (0 \quad 0. \quad 0 \quad 0 \quad 2). \end{array}$$

In this case x_A has 4 significant digits. Note that, if x_A has p significant digits relative to x_T, then 10^{-p} is about the same size as $\texttt{relerr}(x_A)$.

Sometimes, given a number x, we treat it as an x_A, even when we don't have a corresponding x_T. Here are guidelines for assigning the number of significant figures (digits) in a "by itself" x.

- All non-zero digits are considered significant, so 143.2 has four significant figures, while 3.1843 has five significant figures.
- Zeros which lie between non-zero digits are significant, so 1004.3 has five significant figures, and 1324.021 has seven significant figures.
- Zeros which come *before* the first non-zero digit are NOT considered significant, so 0.00001 has only one significant figure, and 00132 has only there significant figures.
- (Hazy rule) Zeros which come *after* the last non-zero digit are considered significant only if indicative of the precision with which the number is known. The number 1000 probably has 1 significant figure, unless, it stems from a scenario such as the following. A meteorite is weighed in kilograms on an accurate scale, and the result is 1000kg, in between 999.0kg and 1001kg. Then 1000 would have 4 significant figures. Typically, for a number like 45.032300, we presume that there are eight significant figures. That is, the last two 0's were reported for a reason, or else the number would have been written as 45.0323.

Loss of significance

Consider two numbers $x = 123.01$ and $y = 123.02$, each known to 5 significant figures. These could be approximations to "true values" $x_T = 123.01289764\cdots$ and $y_T = 123.02175347\cdots$, in which case we might consider them as x_A and y_A. Clearly, the true difference $y_T - x_T = 0.00885582\cdots$ is then a

number with an infinite number of significant digits; it's an exact or pure number. However, working with just x and y, approximations of the true numbers, we see that $y-x=0.01$ has only one significant figure. **The moral of the story:** *subtraction of nearly equal numbers is prone to loss of significance.* When performing numerical (computer) calculations, beware of this trouble.

As an example of a problem where trouble of this stripe is an issue, consider computation of the roots of the quadratic equation $2.1x^2 - 4.5x + 10^{-11} = 0$. OCTAVE's `roots` command yields the following.

```
octave:1> format long e
octave:2> a=2.1; b=-4.5; c=1e-11;
octave:3> rOctave = roots([a b c]);
octave:4> rOctave(1)
ans =    2.142857142854921e+00
octave:5> rOctave(2)
ans =    2.222222222224527e-12
```

These roots are accurate to full double precision (the people who created MATLAB and OCTAVE knew what they were doing)! Let's see what happens when we instead use the straightforward quadratic formula to find the roots.

```
octave:6> r1 = (-b + sqrt(b^2 - 4*a*c))/(2*a)
r1 =    2.142857142854921e+00
octave:7> r2 = (-b - sqrt(b^2 - 4*a*c))/(2*a)
r2 =    2.222137817668790e-12
```

The first expression $r_1 = (-b+\sqrt{b^2-4ac})/(2a)$ is good, and it agrees with OCTAVE's answer r_1^{OCTAVE} from the `roots` command to all digits. However, for the second expression $r_2 = (-b-\sqrt{b^2-4ac})/(2a)$

$$\frac{|r_2 - r_2^{\text{OCTAVE}}|}{|r_2^{\text{OCTAVE}}|} = \frac{|\texttt{2.222137817668790e-12} - \texttt{2.222222222224527e-12}|}{|\texttt{2.222222222224527e-12}|} \simeq \texttt{3.7982e-05}.$$

Viewing r_2^{OCTAVE} as the true root, as an approximation r_2 is only good to a relative error of about `4e-05`, well short of the `1e-16` we'd expect of double precision. Moreover, r_2 has only 4 significant digits relative to r_2^{OCTAVE}.

The bad result for r_2 is not due to the quadratic formula somehow not being valid here; it's a fine formula. If the computer represented numbers exactly, and also performed all arithmetic operations (including square roots) exactly, then it would return an r_2 from the quadratic formula which was perfect. The bad result stems from *computer round-off errors* which lead to loss on significance when subtracting nearly equal numbers. Indeed, consider

```
octave:8> -b
ans = 4.5
octave:9> sqrt(b^2-4*a*c)  % Returns an approximation to the true sqrt(b^2-4ac)
ans = 4.499999999990667
```

So, in forming $-b-\sqrt{b^2-4ac}$ numerically, we are subtracting nearly equal numbers. Danger!

How can we fix and further understand this problem? Here is one easy way. Notice that the "bad" root can be written as

$$\begin{aligned} r_2 &= \frac{-b-\sqrt{b^2-4ac}}{2a} \\ &= \frac{-b-\sqrt{b^2-4ac}}{2a} \overbrace{\left(\frac{-b+\sqrt{b^2-4ac}}{-b+\sqrt{b^2-4ac}}\right)}^{1} \\ &= \frac{2c}{-b+\sqrt{b^2-4ac}} \end{aligned}$$

The denominator of the last expression involves *addition* of nearly equal numbers; there's no problem with that. Let's see what we get using this expression.

```
octave:10> 2*c/(-b + sqrt(b^2-4*a*c))
ans = 2.222222222224527e-12
```

This agrees with r_2^{Octave} to all digits.

Another way to analyze the second root relies on the binomial expansion:

$$\sqrt{1+\delta} = 1 + \tfrac{1}{2}\delta + O(\delta^2),$$

showing $\sqrt{1+\delta} \simeq 1 + \frac{1}{2}\delta$ if $|\delta|$ is small. Consider then

$$\begin{aligned}\sqrt{b^2-4ac} &= \sqrt{(4.5)^2 - 4(2.1)(10^{-11})} \\ &= \sqrt{(4.5)^2 - 8.4\times 10^{-11}} \\ &= 4.5\sqrt{1 - (8.4/4.5^2)\times 10^{-11}},\end{aligned}$$

Set $\delta = -(8.4/4.5^2)\times 10^{-11} \simeq 4.15\times 10^{-12}$ which is small. Then δ^2 is about 1.7×10^{-23}. Then, using the binomial expansion,

$$\sqrt{b^2-4ac} = 4.5\sqrt{1+\delta} \simeq 4.5\big(1+\tfrac{1}{2}\delta\big) = 4.5 - (4.2/4.5)\times 10^{-11} = 4.5 - \tfrac{42}{45}\times 10^{-11}.$$

Using this approximation in the second root, we find

$$\begin{aligned} r_2 &= (2a)^{-1}(-b - \sqrt{b^2-4ac}) \\ &= (4.2)^{-1}(4.5 - \sqrt{b^2-4ac}) \\ &\simeq (4.2)^{-1} \times \tfrac{42}{45} \times 10^{-11}.\end{aligned}$$

This result is then

```
octave:11> (1/4.2)*(42/45)*1e-11
ans = 2.222222222222222e-12
```

This value has 12 significant digits relative to r_2^{Octave}.