Ryan Shìjié Dù
October 21st, 2022

Worksheet 6

Numerical Analysis

The examples in this worksheet comes from [BF10, §1.2].

1 Finite digit arithmetic

For convenience of understanding and calculation, in this worksheet, we model the round-off error as a finite digit arithmetic. We will use k-digit chopping for numbers. For example, if we use 3-digit chopping, and write the finite digit representation as a function $fl()$, we then have

$$\begin{aligned} fl(\pi) &= 3.14 \\ fl(1239.6) &= 1230. \end{aligned}$$

We could also use k-digit rounding where we round the last number instead of chop.

(1.a) Show that for a k-digit chopping representation of numbers, we have

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

2 Quadratic formula: an example of catastrophic cancellation

For the quadratic problem ($a \neq 0$)

$$ax^2 + bx + c = 0$$

the quadratic formula gives the roots:

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \qquad x_x = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Apply this formula to $x^2 + 62.10x + 1 = 0$ whose roots are approximately

$$x_1 \approx -0.01610723, \qquad x_2 \approx -62.08390.$$

(2.a) Use four-digit rounding arithmetic in the calculation to determine the root. Compute the relative error for calculating x_1

$$\frac{|fl(x_1) - x_1|}{|x_1|}.$$

You will find the the relative error is large. Why is this the case?

(2.b) A similar calculation of x_2 using the quadratic formula produces result with a small relative error.

(2.c) To produce a better approximation of x_1, we change the formula by "rationalizing the numerator":

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

Show that this is an equivalent formula as the quadratic formula, for exact arithmetic.

(2.d) Use the new formula to calculate a new $fl(x_1)$, and calculate the relative error. Compare this with the one from using the quadratic formula, why has there been much improvement?

3 Horner's rule for evaluating polynomials

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

(3.a) The most obvious way is to evaluate each term separately, and sum them up. What is the relative error? How many floating point operations are needed (count sum and multiplications separately)?

(3.b) An alternative approach is using the Horner's rule to write the polynomial as a nested expression:

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

Use three-digit chopping arithmetic to evaluate the function. How about the relative error now? How many floating point operations are needed?

References

[BF10] Richard L. Burden and J. Douglas Faires. *Numerical Analysis.* Cengage Learning, August 2010.