Catastrophic Cancellation in log(1+x)

Tae Eun Kim

September 23, 2021

This note explains the catastrophic cancellation observed in Problem 4 of Homework 4.

The evaluation of $f(x)$ is severely affected by catastrophic cancellation for small x because of what is written at the beginning of the problem. Though identical to $f(x)$ mathematically, the function $f_1(x)$ does a better job because of the following reason.

Let $\hat{x} = \widehat{(1+x)} - 1 = fl((1+x)-1)$, the floating-point representation of the expression $(1+x)-1$. Note that the subtraction undergoes *catastrophic cancellation* for small x. Also note that when $\log(1+x)$ is evaluated in the computer, the input $(1+x)$ is formed first and then 1 is subtracted off from it before it is used in the algorithm based on the Taylor series

$$\log \xi = (\xi - 1) - \frac{1}{2}(\xi - 1)^2 + \frac{1}{3}(\xi - 1)^3 - \cdots .$$

Therefore the numerical evaluation of $f_1(x)$ can be approximated by

$$\widehat{f_1(x)} \approx \frac{\hat{x} - \frac{1}{2}\hat{x}^2 + \frac{1}{3}\hat{x}^3 - \cdots}{\hat{x}} = 1 - \frac{1}{2}\hat{x} + \frac{1}{3}\hat{x}^2 - \cdots ,$$

which resembles the series expansion used in part (a). It is clear from the right-hand side that this implementation does not involve subtraction of two nearby numbers for *reasonably* small x. However, when x is sufficiently small, $(1+x)$ is indistinguishable from 1 on the floating-point number system, in which case

$$\hat{x} = \widehat{(1+x)} - 1 = 0.$$

We know that it happens when x is smaller than the machine epsilon `eps`, which is about 2×10^{-16}; in our experiment, it happens for $k \geq 16$, in which case both the numerator and the denominator of $f_1(x)$ are evaluated as zeros, resulting in `NaN`.

One way to avoid the issue with `NaN` is to use the series expansion obtained in part (a), namely,

$$f(x) = 1 - \frac{1}{2}x + \frac{1}{3}x^2 - \cdots .$$

Another way is to use the function `log1p` as suggested in the problem. This function was specifically designed to avoid catastrophic cancellation occurring in the evaluation of $\log(1+x)$ for small x by encoding

$$\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots$$

instead of using the series expansion for $\log \xi$ written above.