Assignment 1 - bridging course

Linus Falk

30 oktober 2022

1 Workout 2.9

Prove that if x and y are floating-point numbers with $\frac{y}{2} \leq x \leq y$ then $\mathbf{fl}(x-y) = x - y$ provided that the guard digit is supported, $\beta = 2$ and $x - y$ does not underflow.
Let $y = (d_0 d_1 d_2 \dots d_{p-1})\beta^e$ Then we have:

$$(d_0 d_1 d_2 \dots d_{p-1})\beta^{e-1} \leq x \leq (d_0' d_1' d_2' \dots d_{p-1}')\beta^e \tag{1}$$

if we investigate the inequality we can see that x is between the exponent e and $e-1$. For every x with the same e as y we have exact subtraction of the two floating-type numbers. In the other case we have a maximum difference of 1 in the exponent. If swap between x and y, such that $0 \leq y \leq x$ and scale them so we can represent the digits in x as in (2) $(x_0 x_1 \dots x_{p-1})$ we will have exact subtraction because $x \leq y$ and $x - y \leq y$, so we will only have p valid digits and $w_0 = 0$

$$\begin{array}{cccccc} x_0 & x_1 & x_2 & \dots & \dots & x_{p-1} \\ 0 & y_0 & y_1 & \dots & \dots & y_p \\ \hline w_0 & w_1 & w_2 & \dots & \dots & w_p \end{array} \tag{2}$$

2 Workout 2.10

Consider the Newton's method for solving $f(x) = a - \frac{1}{x}$ for a given real number a. Show that the computation of $x_{k+1} from x_k$ (successive Newton's iterations) can be expressed as two multiply-adds, thus the roundoff errors is reduced by a factor of $\frac{1}{4}$ if FMA operation is available.

2.1 Newtons method

$$x_{k+1} = x_k - \frac{f(x)}{f'(x)} \tag{3}$$

We find the derivative of $a - \frac{1}{x}$ that is $\frac{1}{x^2}$ and plug into (2) and simplify to see if we can put it in two multiply-adds form.

$$x_{k+1} = x_k - \frac{a - \frac{1}{x_k}}{\frac{1}{x^2}} = x_k - x_k^2 \left(a - \frac{1}{x_k}\right) = x_k - x_k \left(ax_k - 1\right) \tag{4}$$

By putting the equation on the later form we can compute the iterations with two multiply-adds. First inside the parenthesis and then the outer part.