

Let us use a decimal floating point system  $F_0$  with 4 digits and an exponent between -9 and 9.

$$\{\beta = 10, t = 4, u = -L = 9\}$$

(Associativity)

Take  $a = 1.000 \cdot 10^0$ ,  $b = 3.000 \cdot 10^{-4}$ ,  $c = 4.000 \cdot 10^{-4}$

$$a + b = 1.000 \cdot 10^0 + 0.0003 \cdot 10^0 = 1.000 \cdot 10^0$$

$$(a + b) + c = 1.000 \cdot 10^0 + 0.0004 \cdot 10^0 = 1.000 \cdot 10^0$$

$$b + c = 3.000 \cdot 10^{-4} + 4.000 \cdot 10^{-4} = 7.000 \cdot 10^{-4}$$

$$a + (b + c) = 1.000 \cdot 10^0 + 0.0007 \cdot 10^0 = 1.001 \cdot 10^0$$

Not associative!

Uniqueness of solutions to algebraic equations does not hold for floating point arithmetic.

$1 + x = 1$  has more than 1 solution in  $F$  not just the "exact" solution  $x = 0$ .

In fact in  $F_0$ , every positive floating point number  $y < 5.000 \cdot 10^{-4}$  is a correct solution to the equation.

In the course of solving  $ax^2 - 2bx + c = 0$ , for  $x$ , the expression  $\sqrt{b^2 - ac}$  must be computed. Can the true value of  $b^2 - ac$  be non-negative and yet have a negative computed value?

$$f(b) = b(1 + \delta_1) \text{ with } |\delta_1| < u \text{ the unit roundoff}$$

$$\text{Then } f(b^2) = (b(1 + \delta_1))^2(1 + \delta_2) \text{ with } |\delta_2| < u$$

$$f(a) = a(1 + \eta_1) \text{ with } |\eta_1| < u$$

$$f(c) = c(1 + \eta_2) \text{ with } |\eta_2| < u$$

$$f(ac) = ac(1 + \eta_1)(1 + \eta_2)(1 + \eta_3) \text{ with } |\eta_3| < u$$

$$f(f(b^2) - f(a)f(c)) = [b^2(1 + \delta_1)^2(1 + \delta_2) - ac(1 + \eta_1)(1 + \eta_2)(1 + \eta_3)](1 + \epsilon_1) \text{ where}$$

$$|\epsilon_1| < u$$

$$\approx b^2 - ac + b^2(2\delta_1 + \delta_2) - ac(\eta_1 + \eta_2 + \eta_3) + \epsilon_1(b^2 - ac)$$

and products & squares of small quantities have been neglected.

Take

$$a = 1.998$$

$$b = 1.003$$

$$c = 0.503$$

In three digit decimal point arithmetic.

$$f(f(b^2) - f(a)f(c)) = (1.00)^2 - (2.00)(0.503) = -0.01 < 0$$

The quadratic formula states that the roots of  $ax^2 + bx + c = 0$ , when  $a \neq 0$ , are

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

Consider  $x^2 + 62.10x + 1 = 0$  with roots  $x_1 = -0.01610723$ ,  $x_2 = -62.08390$

Note that  $b^2$  is much larger than  $4ac$  so the numerator in  $x_1$  involves subtraction of nearly equal numbers.

Lets use our  $F_0$  in 4 digit arithmetic.

$$f(\sqrt{b^2 - 4ac}) = f(\sqrt{62.10^2 - 4.000}) = \sqrt{3856 - 4.000} = 62.06$$

$$f(x_1) = \frac{-62.10 + 62.06}{2.000} = -\frac{0.0400}{2.000} = -0.020$$

Relative error in  $x_1$

$$\frac{|-0.0161 + 0.010|}{|-0.0161|} = 2.4 \cdot 10^{-1}$$

For the other root, no cancellation;

$$f(x_2) = \frac{-62.10 - 62.06}{2.000} = \frac{-124.2}{2.000} = -62.10$$

So relative error in  $x_2$

$$\frac{|-62.08 + 62.10|}{|-62.08|} \approx 3.2 \cdot 10^{-4}$$

To obtain a more accurate four-digit rounding approximation to  $x_1$ , we change the form of the quadratic formula, by rationalising the numerator.

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \cdot \frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}}$$

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

$$f(x_1) = -\frac{2.000}{62.10 + 62.06} = -\frac{2.000}{124.2} = -0.01610$$

With relative error

$$\frac{|-0.01611 + 0.01610|}{|-0.01611|} \approx 6.2 \cdot 10^{-4}$$

So we could also use rationalisation on  $x_2$

$$x_2 = \frac{-2c}{b - \sqrt{b^2 - 4ac}}$$

$$f(x_2) = -\frac{2.000}{62.10 - 62.06} = -\frac{2.000}{0.0400} \approx -50.00$$

$\Rightarrow$  a large relative error of  $1.9 \cdot 10^{-1}$