

The problem of interest consists in finding the smallest solution x_1 of the equation $x^2 - 2px + q$ under the assumption $p, q > 0$ and $p^2 > q$. In particular we are interested in the situation $p^2 \gg q$.

Of course we have an explicit solving formula for this problem, namely

$$x_1 = p - \sqrt{p^2 - q} =: f(p, q).$$

It allows us to compute the condition number of the problem with the formula

$$K_f \approx \left| \frac{p \frac{\partial f}{\partial p}}{f} \right| + \left| \frac{q \frac{\partial f}{\partial q}}{f} \right| =: A + B$$

$$\frac{\partial f}{\partial p} = 1 - \frac{p}{\sqrt{p^2 - q}} = -\frac{p - \sqrt{p^2 - q}}{\sqrt{p^2 - q}}$$

so that

$$A = \frac{p}{\sqrt{p^2 - q}} = \frac{1}{\sqrt{1 - q/p^2}}$$

On the other hand

$$\frac{\partial f}{\partial q} = -\frac{-1}{2\sqrt{p^2 - q}}$$

so that, multiplying and dividing by $p + \sqrt{p^2 - q}$ and then simplifying q

$$B = \left| \frac{q}{2(p - \sqrt{p^2 - q})\sqrt{p^2 - q}} \right| = \frac{1 + \sqrt{1 - q/p^2}}{2\sqrt{1 - q/p^2}}$$

Using that $0 < \sqrt{1 - q/p^2} < 1$, we obtain the bounds

$$\frac{1}{2\sqrt{1 - q/p^2}} < B < \frac{1}{\sqrt{1 - q/p^2}}$$

Using the upper bound, we obtain:

$$A + B < \frac{2}{\sqrt{1 - q/p^2}}.$$

This can grow to infinity. This happens when $p^2 \approx q$, the case of two almost coincident roots. However if $p^2 \gg q$, we have $q/p^2 \ll 1$ and $\sqrt{1 - q/p^2} \approx 1$, giving a condition number $K \approx A + B < 2$: the problem is well-conditioned.

The obvious algorithm to compute x_1 is given by

$$\text{flt}(p - \sqrt{p^2 - q}), \tag{1}$$

however the external subtraction (last nontrivial residual transformation) incurs in a cancellation error, if $p^2 \gg q$, since $p \approx \sqrt{p^2 - q}$. The rounding error generated e.g. by the square root (ϵ_M) will be amplified by a factor

$$K_- \approx \frac{p + \sqrt{p^2 - q}}{p - \sqrt{p^2 - q}} \approx \frac{2}{1 - \sqrt{1 - q/p^2}}$$

Denoting by $s := q/p^2$ (a small quantity, if $p^2 \gg q$) we can Taylor-expand the square root as $\sqrt{1 - s} = 1 - \frac{1}{2}s + \mathcal{O}(s^2)$ to obtain

$$K_- \approx \frac{4}{s}.$$

As an example, if $p \approx 1/2 \cdot 10^8$ and $q \approx 1/3 \cdot 10^8$ (values chosen such that the solutions are $x_1 = 1/3$ and $x_2 = 10^8$) we have $s \approx 4/3 \cdot 10^{-8}$ and the condition number of the residual transformation is of the order of magnitude of 10^8 , meaning that we lose (in a single floating point operation) eight significant digits in base 10. The alternative algorithm

$$\text{flt} \left(\frac{q}{p + \sqrt{p^2 - q}} \right)$$

is on the contrary stable. This is a direct consequence of the fact that all involved elementary operations are well conditioned (in the regime $p^2 \gg q$), so that also the residual transformations (obtained by composition) are well conditioned.