Error Analysis

A fundamental question for any computation is "how good is my answer?" Error analysis is an analytical attempt to address this question. There are two different perspectives that can be taken with such an analysis: forward error and backward error. We will show each perspective by analyzing the rounding errors in the addition of two real numbers. The floating point representation theorem (FRT) and the fundamental axiom of floating point arithmetic (FAFA) will be our tools.

Our forward error analysis will find an upper bound on the rounding errors in the addition of $x, y \in \mathbb{R}$. We assume $x,\ y$ and $x + y$ are in the range of the floating point system. Can we find an upper bound on the relative difference between $x + y$ and $\mathrm{fl}(\mathrm{fl}(x) + \mathrm{fl}(y))$? By the FRT there exist $|\delta_x|, |\delta_y| \leq \boldsymbol{\mu}$, and by the the FAFA $|\delta| \leq \boldsymbol{\mu}$ such that (working from the inside out)

$$\begin{aligned} \mathrm{fl}(\mathrm{fl}(x) + \mathrm{fl}(y)) &= \mathrm{fl}(x(1+\delta_x) + y(1+\delta_y)) \\ &= [x(1+\delta_x) + y(1+\delta_y)](1+\delta) \\ &= x(1+\delta_x+\delta+\delta_x\delta) + y(1+\delta_y+\delta+\delta_y\delta) \\ &= x + y + x(\delta_x+\delta) + y(\delta_y+\delta) + x(\delta_x\delta) + y(\delta_y\delta) \end{aligned}$$

Then

$$\begin{aligned} |\mathrm{fl}(\mathrm{fl}(x) + \mathrm{fl}(y)) - (x+y)| &\leq |x(\delta_x+\delta)| + |y(\delta_y+\delta)| + |x(\delta_x\delta)| + |y(\delta_y\delta)| \\ &\leq 2\boldsymbol{\mu}(|x| + |y|) + \boldsymbol{\mu}^2(|x| + |y|). \end{aligned}$$

A relative error bound is therefore

$$\frac{|\mathrm{fl}(\mathrm{fl}(x) + \mathrm{fl}(y)) - (x+y)|}{|x+y|} \leq 2\boldsymbol{\mu}\frac{|x| + |y|}{|x+y|} + \mathrm{O}(\boldsymbol{\mu}^2).$$

When is this upper bound small? When is it large?

Our backward error analysis will find a bound on ϵ_x and ϵ_y such that $\mathrm{fl}(\mathrm{fl}(x) + \mathrm{fl}(y)) = (x + \epsilon_x) + (y + \epsilon_y)$:

$$\begin{aligned} \mathrm{fl}(\mathrm{fl}(x) + \mathrm{fl}(y)) &= \mathrm{fl}(x(1+\delta_x) + y(1+\delta_y)) \\ &= [x(1+\delta_x) + y(1+\delta_y)](1+\delta)\ , \\ &= (x + x\epsilon_x) + (y + y\epsilon_y) \end{aligned}$$

where $\epsilon_x = \delta_x + \delta + \delta_x\delta$ and $\epsilon_y = \delta_y + \delta + \delta_y\delta$. Then

$$\left|\frac{x\epsilon_x}{x}\right|, \left|\frac{y\epsilon_y}{y}\right| \leq 2\boldsymbol{\mu} + \boldsymbol{\mu}^2.$$

This result says that we have computed the exact sum of two numbers relatively close to x and y. A backward error result does not give us an upper bound on the error; it gives us an upper bound on the difference between the problem we wanted to solve and a problem that we did solve. Unlikely as it may seem in this simple example, it is backward error analysis that is more useful in practice.